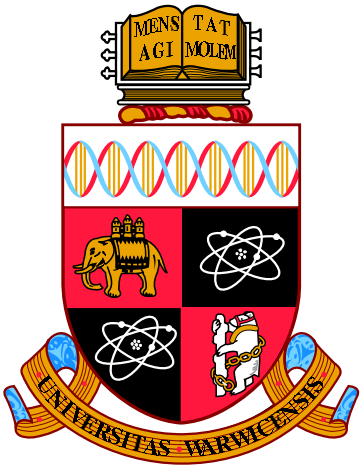


Machine Learning Stratification for Oncology Patient Survival

by

Katherine Louise Lloyd

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Supervisors: Dr. Richard S. Savage, Prof. Ian A. Cree

MOAC Doctoral Training Centre

September 2017

THE UNIVERSITY OF WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

I would like to thank everyone who contributed in some way to the work described in my thesis.

First and foremost, I would like to thank my supervisors, Dr. Richard Savage and Prof. Ian Cree, for their advice, encouragement, time, and expertise. I am extremely grateful for all their support, both academic and personal.

I would also like to thank my advisory committee members, Dr. David Snead, Prof. Janet Dunn, and Prof. David Wild for their support and guidance, and their challenging questions.

Many thanks to all the staff and students of MOAC DTC over the past five years; the experience was much enriched by the sense of community I found here. Also, many thanks to Team Savage, for their upbeat and encouraging discussions.

I also thank the staff and students I met and worked with at UHCW, with thanks to Hugh Kikuchi, Tina Wotherspoon and Jenifer Nyoni for their patience around the lab. I would also like to extend my thanks to Anne Reiman, for her support, patience and guidance in the lab, and in life in general.

I would like to extend special thanks to Hugo van den Berg for his constructive criticism and his much appreciated support during trying circumstances.

For their friendship and unending support, I would like to say a big thank you to Haze, Faz and Ruth.

Finally, as always, I would like to extend my sincere gratitude to my family, and in particular my parents George and Fran, who helped me though this challenging but highly rewarding experience.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. The work presented (including data generated and data analysis) was carried out by the author.

Parts of this thesis have been published in or submitted to scientific journals by the author.

Chapter 2

Katherine L Lloyd, Ian A Cree, and Richard S Savage. Prediction of resistance to chemotherapy in ovarian cancer: a systematic review. *BMC cancer*, 15(1): 117, 2015.

Chapter 3 and **Chapter 4**

Katherine L Lloyd and Richard S Savage. Gaussian processes for survival data with right-censoring. **submitted**.

Abstract

Personalised medicine for cancer treatment promises benefits for patient survival and effective use of medical resources. This goal requires the development of predictive models for the identification and implementation of biomarkers for the prediction of patient survival given treatment options. This thesis addresses research questions in this area.

The systematic review detailed in Chapter 2 investigates the literature concerning the prediction of resistance to chemotherapy for ovarian cancer patients using statistical methods and gene expression measurements. The range of models used by studies in the systematic review highlights the popularity of traditional models, such as Cox proportional hazards, with few more complex models being utilised.

In Chapters 3 and 4, new methods are presented for modelling right-censored survival data. Due to the nature of biomedical data, the methods used need to be flexible and adequately account for high dimensional, noisy data. Gaussian processes fulfil these requirements and were hence used for the development of three Gaussian process models for right-censored survival data. Chapter 3 details these models, and they are applied to synthetic and cancer data in Chapter 4. In all cases the Gaussian processes for survival were found to equal or outperform all comparison models, as measured by concordance index.

Given the application to molecular cancer data, it was expected that the data would be high dimensional. Two feature selection methods are investigated in Chapter 5 for use with Gaussian processes to address this.

In Chapter 6 a program is developed for the analysis of the data produced by a test for cancer mutations using qPCR. The automated program was designed to isolate the analysis from the user and produce results and reports for clinical use. It is observed that this approach of automated analysis would be suitable for application to any form of clinical test or complex predictive model without the requirement of user guidance.

Abbreviations

- AFT: Accelerated failure time
- AIC: Akaike information criterion
- ANOVA: Analysis of variance
- ARD: Automatic relevance determination
- AUC: Area under the curve
- BIC: Bayesian information criterion
- BMA: Bayesian model averaging
- CEA: Carcinoembryonic antigen
- CEBM: Centre for Evidence-Based Medicine
- CI: Confidence interval
- COSMIC: Catalogue Of Somatic Mutations In Cancer
- CoV: Coefficient of variance
- Dim: Dimension
- DNA: Deoxyribonucleic acid
- FFPE: Formalin fixed paraffin embedded
- GBM: Gradient boosting machine

- GBMF: Glioblastoma multiforme
- GP: Gaussian process
- GPR: Gaussian process regression
- GPS: Gaussian process for survival data
- GSEA: Gene set enrichment analysis
- HR: Hazard ratio
- HUGO: Human Genome Organisation
- IARD: Informed automatic relevance determination
- ISIS: Identifying splits with clear separation
- KEGG: Kyoto Encyclopedia of Genes and Genomes
- KIRC: Kidney renal clear cell carcinoma
- LR: Likelihood ratio
- LUSC: Lung squamous cell carcinoma
- miRNA: Micro ribonucleic acid
- mRNA: Messenger ribonucleic acid
- NA: Not applicable
- NHS: National Heath Service
- NICE: National Institute for Health and Care Excellence
- NPV: Negative predictive value
- NS: Not specified
- OCGS: Ovarian Cancer Gene Set

- ONS: Office of National Statistics
- OS: Overall survival
- OV: Ovarian serous cystadenocarcinoma
- PFS: Progression-free survival
- PPV: Positive predictive value
- PSA: Prostate specific antigen
- qPCR: Quantitative polymerase chain reaction
- QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2
- RF: Random Forest
- RFS: Relapse-free survival
- ROC: Receiver-operator curve
- RSF: Random Survival Forest
- RSFS: Random subset feature selection
- RT-PCR: Reverse transcription polymerase chain reaction
- SCNA: Somatic copy-number alteration
- SqExp: Squared exponential
- SRGS: Systematic Review Gene Set
- TCGA: The Cancer Genome Atlas
- WT: Wild type

Chapter 1

Introduction

The thesis presented here addresses research questions relating to personalised medicine for cancer and the application of tools to predict patient survival in a clinical setting. Gaussian process models developed to infer survival outcomes are introduced, implemented and tested. These models are investigated in the context of oncology patient survival, using molecular and clinical data. Also introduced are two feature selection techniques applied to Gaussian processes for use with these high-dimensional data. Additionally, a published systematic review is presented, investigating the prediction of treatment response in ovarian cancer, and a program for the analysis of qPCR-based mutation testing data for clinical use is detailed.

1.1 Oncogenesis

The average rate of cancer incidence in England was 595.8 per 100 000 people in 2015 [44], with an average 5-year age-standardised mortality rate of 59% for patients diagnosed between 2010 and 2014 and followed up to 2015 [10]. In the Cancer Registration Statistics bulletin [44] released by the Office for National Statistics in 2017, it was observed that although on average 'cancer mortality rates have generally decreased over time, despite the increase in cancer incidence', this trend is not observed across all cancer types. For example, whereas breast and prostate cancer mortality rates have fallen, this trend is not found for lung cancer in women, pancreas and larynx in men [2]. For breast and prostate cancer, increase in incidence and decrease in mortality both may be attributed in part to improvement in diagnostic tools: the national breast screening program for women aged 50 to 70, and more widespread use of the prostate-specific antigen test in older men [43]. Additionally, it is observed that response to breast cancer treatment may be dependent on ER

and HER2 status, and treatments are recommended to differ based on this [112]. It is therefore clear that the development of diagnostic and treatment-specific tools is of great importance to the decrease in cancer mortality rates.

Cancer is characterised by uncontrolled replication of cells [82]. In a 'normal' cell, there are a range of processes that monitor and control the vital cellular mechanisms such as growth, division, and metabolism. These control processes prevent the cell from unconstrained responses that, whilst beneficial for the individual cell and its descendents, may be deleterious to the organism as a whole. When these processes are hindered, therefore, the cell may undergo unrestrained growth, for example [159]. It is interesting that, given deactivation of genetic stability control processes, mutations become more common, often resulting in the deactivation of further processes [93].

The causes of cancer are still a highly active research area, but it is generally accepted that DNA mutations play a key role in oncogenesis [23]. Cellular proteins are produced from amino acids and peptides using DNA as a template. Any damage in this template, therefore, will result in changes in the proteins produced. Depending on the magnitude of the change, the protein may become non-functional, or the function may change. It is thought that, due to mutations, cancerous cells have DNA that does not code for the original proteins, causing alterations in cellular processes and changes in rates of cell growth, replication, and death. For example, the loss of function of proteins involved in important monitoring processes may allow cells to grow and replicate uncontrollably.

There are types of genes, known as oncogenes and tumour-suppressor genes, that are particularly important in oncogenesis [82]. Oncogenes are identified as those that, when activated, promote cancer. These genes are usually inactive or active at a low level but when they are activated, for example via a change of function mutation or somatic copy number mutation, they produce cancer-promoting products or overexpress normal products. Many oncogenes, for example RAS [95], start as proto-oncogenes and perform a cellular function before activation to become an oncogene. Tumour suppressor genes code for proteins involved in inhibitory pathways that control the cell cycle or apoptosis. Mutations in these genes may lead to loss of function in proteins and hence lack of regulation. For example, p53, a gene coding for a protein strongly linked to cell cycle arrest and apoptosis, conveys protection against the proliferation of cancerous cells and is found to be mutated in a large proportion of tumour types [143].

The genetic locations of any mutations present in the genome are therefore of interest, as knowledge of the 'turned off' processes may allow patterns in tumour

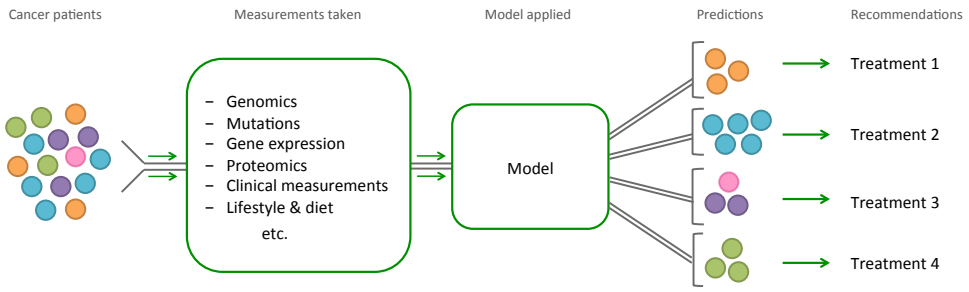


Figure 1.1: An outline of personalised medicine. The patient population is expected to have varying disease characteristics. For each patient, disease and patient-specific measurements are taken and passed to the model. The model makes predictions which may be used to make actionable recommendations for treatment guidance.

development to be predicted.

1.2 Personalised medicine for cancer

Cancer is highly heterogeneous, both in development and treatment response [97]. This is partially due to the high variation in mutations present in different cancers, and even within tumours, and presents challenges for treatment [97]. Personalised medicine proposes that, following testing, treatments should be selected based on tumour-specific characteristics [65]. Currently, the NHS is starting to phase in mutation testing for a small range of cancers, for example non-small-cell lung cancer [108, 106, 111] and metastatic melanoma [110, 109], but this is limited in terms of both tests and the available treatment choices. However, the NHS Personalised Medicine Strategy is aiming for personalised medicine in cancer to be started to be implemented by 2020 [1], presenting the need for tests and analysis tools capable of much higher throughput and complex analysis. Figure 1.1 shows a diagram depicting the application of personalised medicine.

Prognosis is defined as the 'likely outcome or course of a disease; the chance of recovery or recurrence' [5]. Prognostic models and biomarkers are therefore those that predict patient survival using relevant clinical and non-clinical information. These models, however, do not take treatments and interventions into account. Models and biomarkers predicting survival, given a certain therapy are referred to as predictive. Prognostic models are a special case of predictive models, given no treatment.

For example, consider two biomarkers: prostate specific antigen (PSA) as a biomarker for prostate cancer survival, and EGFR mutation status for survival following erlotinib maintenance treatment for advanced non–small-cell lung cancer

[13]. In the first case, PSA expression is used as a prognostic biomarker whereby high expression predicts short survival times, but does not influence treatment choice. By contrast, in the second case the presence or absence of an EGFR mutation is used as a predictive biomarker whereby the presence predicts better response to erlotinib, but the survival estimate is also dependent on the treatment. Considering the discrete case for simplicity, for the prognostic biomarker there are only two possible cases – biomarker positive or biomarker negative – whereas for the predictive biomarker there are four, due to the binary nature of both the biomarker and treatment use. In this way, predictive biomarkers are capable of incorporating more information than prognostic biomarkers when utilised for treatment selection.

Similarly, models may be prognostic or predictive. Clinically, the two model types are often used differently [160]. Prognostic models are often based on non-disease-related information and used at an early time point to predict patient survival or disease progression. Measurements used in these models may include factors such as age or whether a patient smokes. Predictions using these models may provide assistance when considering options such as watchful waiting or surgery [160]. By contrast, predictive models can be used to predict patient survival under the influence of a treatment or intervention. For example, the Mammaprint$^{\text{TM}}$ gene signature [146] uses the expression of 70 genes and HER status to predict breast cancer patient likelihood of metastasis and the utility of chemotherapy.

1.3 Prediction of response to cancer treatment

Resistance to chemotherapy is a well documented response, and treatment plans usually account for this eventuality when recommending therapies [28]. For advanced stage ovarian cancer, for example, NICE recommends platinum-based first-line treatment either alone or in combination with paclitaxel [105]. However, dependent on the response to platinum-based therapy, second line treatments are more varied, as treatment of a platinum-resistant tumour with platinum-based therapy would not be successful.

There are two key proposed mechanisms by which cancers become treatment resistant. First, as therapies are designed to kill tumour cells, there is an element of clonal selection. It is suggested that, given a group of tumour cells with a variety of mutations, those that survive chemotherapy will later form the tumour present at second-line treatment [56]. For example, in BRCA mutated cancer, a second BRCA mutation has been seen to restore BRCA functionality and improve DNA repair, counteracting the mechanism of action of platinum-based treatments [94].

For treatments such as anti-estrogen treatment for breast cancer, loss of estrogen receptors has been observed, removing the drug target and hence preventing cell death [79].

An alternative route of chemotherapy resistance is through changes in gene expression, for example via altered epigenetics, allowing the up and down-regulation of pathways relating to cell survival [56]. There are a range of mechanisms by which protection may be inferred, including increased DNA repair, increased drug pumps and detox, reduction in drug target, decreased apoptosis, and increased proliferation [54].

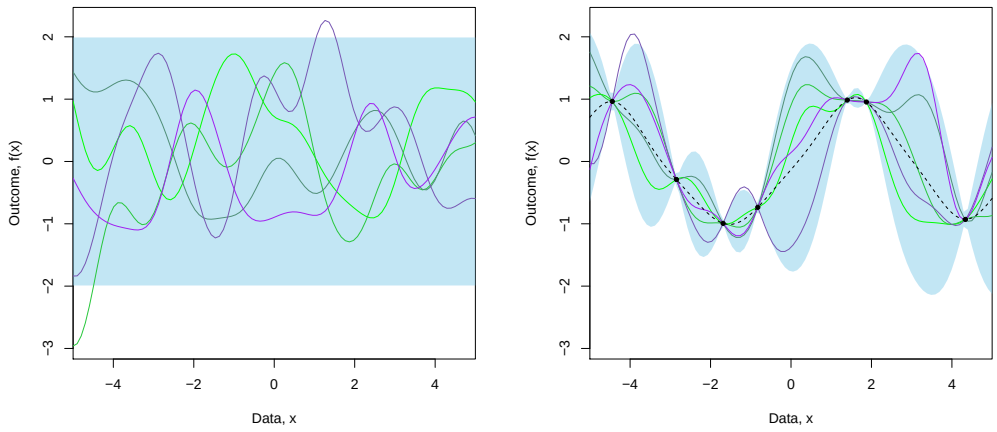
It is expected that knowledge of mutations and gene expression levels may allow the prediction of patient response to chemotherapy. If this is the case, treatment pathways may be personalised and treatments to which patient response would not be optimal may be avoided, greatly improving survival and patient quality of life.

1.4 Statistical modelling and machine learning

Given the volume of data often available about a given cancer patient, there is great potential value in developing tools that can produce clinically actionable recommendations. These tools should be capable of taking in data and producing a personalised recommendation for a patient, enabling clinical decisions to be made. These tools would involve the application of a statistical or mathematical model, allowing underlying trends in the data to be identified and utilised for prediction.

In this context, statistical models may be very useful; given potentially noisy data and assuming a given probability distribution, they may predict likely outcomes [99]. For example, given data and outcomes X and $\mathbf{y}$, a linear regression formulation calculates the optimum model parameters β such that the fit is optimised and hence the relationship is decided to be $\mathbf{y} = X\beta + \epsilon$ [104], where ϵ is Gaussian noise. For this model, the relationship is pre-determined to be linear, errors are assumed to be uncorrelated and normally distributed, as are the covariates, as well as further assumptions.

In contexts with small numbers of variables, and where the underlying data-generating mechanisms are thought to be understood, well-defined models such as linear regression have great utility. However, these models do not generalise well to situations where relationships between variables are not known or are complex. In the context of molecular cancer data and similar biological measurements, models capable of more complex relationships between variables and outcome are therefore required.



(a) (b)

Figure 1.2: a) Functions randomly drawn from a GP prior. b) Functions randomly drawn from a GP posterior, given data X. Data points are marked as dots, the predicted mean is marked as a dotted line. The shaded area represents the 95% confidence interval (2 standard deviations from the predicted mean). Based on Figure 2.2 from Rasmussen and Williams [127].

Many machine learning methods offer flexibility and the ability to identify patterns in data [104]. Additionally, as with some statistical models, the predictions produced by some machine learning methods may be probabilistic. This has the benefit of allowing the confidence in the predicted values to be determined, which would clearly be of use in clinical contexts. Models such as Gaussian processes provide the benefits of both statistical and machine learning techniques, and hence are particularly suitable for use with complex, noisy medical data.

Gaussian processes consider the function $f(\mathbf{x})$ as coming from the space of functions relating features to outcome, $y = f(\mathbf{x}) + \epsilon$, subject to a prior constraining $f(\mathbf{x})$ to have certain qualities, such as smoothness and stationarity [127]. Given the data, they provide a posterior distribution on the possible functions. In this way, the properties of the possible functions may be used to predict the posterior mean and variance for given test $\mathbf{x}_*$ values. For Gaussian processes the prior is determined by the choice of covariance and mean functions, and the associated hyperparameters. Given two data points $\mathbf{x}_1$ and $\mathbf{x}_2$, the covariance and mean functions define how similar the values of the function, $f(\mathbf{x}_1)$ and $f(\mathbf{x}_2)$, must be.

The effect of conditioning the prior using the data X may be seen in Figure 1.2, where functions drawn randomly from the prior and posterior are shown.

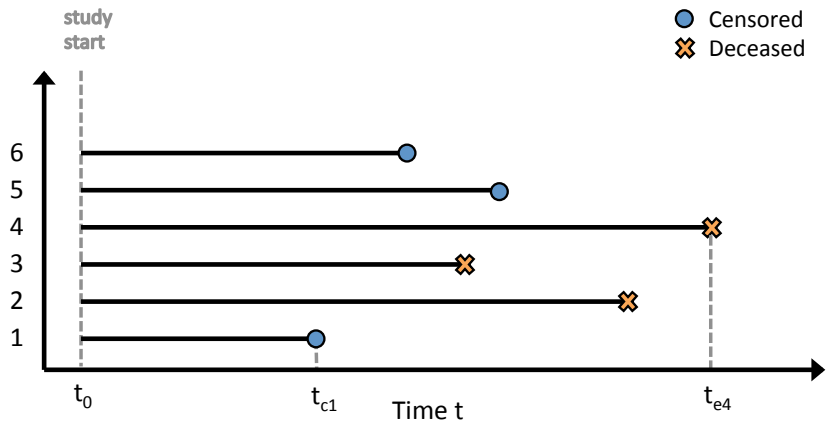


Figure 1.3: Illustration of right-censored survival times. Censored times are marked with dots, uncensored times as crosses. The unknown true survival times for the censored samples will be further to the right than the censored times; the censored times act as minimum values for the true times.

1.5 Modelling and predicting survival outcomes

Medical studies involving patients often produce survival data. Generally, all patients enter the study at a measurable time, t_0, and the required measurements are recorded as the study progresses. The patients are monitored for the occurrence of a pre-defined event, examples of which could be death, disease progression or relapse. If this event occurs, the time at which this happens is recorded for each patient, t_e. These times are referred to as survival times. Traditionally, when incorporated into a model these survival times are given the label $\mathbf{y}$ and are also known as targets. The measured variables, also known as features, are assigned the label X.

It is common in studies involving patients for a number of patients to drop out before the end of the study. This may be due to a variety of reasons and results in the patient being lost to follow-up for the remainder of the study time. Data X for these patients will normally have already been collected, so the data are not discarded; the survival time is instead considered to be right-censored. A right-censored survival time is the time at which the patient was last seen, and provides a lower bound for the unknown date of event occurrence, $t_c \leq t_e$. Censored survival data are depicted in Figure 1.3.

For a data set containing multiple patients, survival data therefore consists of survival times $\mathbf{y}$, measured variables X, and whether each patient underwent the event, $\mathbf{e}$.

Currently in the literature there are many commonly used models for survival data. The most popular of these, Cox proportional hazards regression [29], relies on the assumption that the risk of outcome is proportional to the exponential of a linear combination of the covariates [40]. This assumption allows the proportionality constant, the baseline hazard, to be disregarded. This approach is sensible when considering hazard ratios, as the baseline hazard will cancel. However, when attempting to use Cox proportional hazards regression predictively, the baseline hazard may become problematic to estimate. Other popular methods include accelerated failure time models [163] and gradient boosted machines [138].

Whilst these models are well researched and frequently used, in the context of medical molecular data it becomes increasingly necessary for machine learning techniques to be available for survival data. Gaussian processes provide a range of existing, well researched approaches and techniques for modelling data from a range of contexts. Regression and classification are routinely applied for real and categorical outcomes. Due to their flexibility, ability to work with high dimensional data, and inherent ability to account for noisy data, Gaussian processes would be an ideal tool for medical research data analysis. However, Gaussian processes for survival data are a lesser researched topic.

An approach to Gaussian processes for survival data is considered here in Chapter 3. Given the form of right-censored survival data, it is sensible to consider censored times as partially missing outcomes, with the censored time acting as a lower bound on the possible values. Gaussian processes are therefore used to place a prior on the space of functions relating features to outcomes, with the aim of inferring the missing values. This approach attempts to predict the underlying, non-censored outcomes using information from both the uncensored and censored samples.

1.6 Cancer patient data and measurements

As medical data, the measurements relevant in cancer research involve various clinical features. Research is usually separated by the primary tumour site, as it is well observed that cancers in different sites progress very differently [70]. Stage and grade are often reported, and both are used as measures of tumour progression. Cancer stage takes into account tumour size and location, whether lymph nodes are involved, and whether the primary tumour has metastasised [4]. For most cancers, stages range from 1 to 4, with some also having stage 0. Tumour grade is a measure of the level of similarity between the tumour tissue and normal cells for the site, where the normal cells are well-differentiated [6]. It is expected that as the cancer

progresses cells will undergo anaplasia, whereby they revert to an undifferentiated state. Undifferentiated cells are known to grow and divide faster, and hence would lead to faster tumour growth and metastasis [6]. Other measurements of interest include age, previous treatments, and relevant site-specific biomarker levels, such as BRCA1/BRCA2, HER2, CA-125, PSA and CEA.

Mutation analysis is highly relevant in cancer research. This requires a method of sequencing, to identify the order of bases that make up the DNA of the genes of interest. Two common techniques are next generation sequencing, and the quantitative polymerase chain reaction (qPCR). Next generation sequencing provides a platform to generate large amounts of sequence data without requiring prior knowledge of likely mutations. By contrast, qPCR methods are more targeted, with a list of mutations being tested, but are often faster and cheaper to run.

Gene expression is often also of interest in cancer studies and involves measuring the levels of mRNA within cells. Due to the central dogma - DNA is transcribed to mRNA, which is translated to proteins - measuring the levels of mRNA allows the required levels of protein to be inferred [83]. As cancer cells up-regulate processes that aid their survival and down-regulate those that would impose control mechanisms, the levels of expression of a wide range of genes are expected to be abnormal in cancer cells [132]. It is therefore reasonable to expect that knowledge of gene expression may provide information on the status and likely trajectory of cancer development and response to treatment. Gene expression is often measured via qPCR, with common commercial kits providing measurements of around 20 000 genes [36]. These data are therefore very high dimensional.

1.7 Feature selection

Due to the often high-dimensionality of cancer data, feature selection is often applied. This method involves the identification of important and informative features, and the removal of those that are deemed unnecessary.

Although the addition of features would be expected to increase the information available, the inclusion of many dimensions of data into a model may cause computational and statistical issues, due to the 'curse of dimensionality' [17]: as the number of features increases the volume of the feature space grows exponentially, and the samples become increasingly sparse, and often unevenly distributed though the space. Moreover, the inclusion of additional features inherently adds noise, which for features with little predictive value may outweigh the added signal. Therefore, in order to obtain better predictions, it may be useful to remove non-informative

features, reducing the dimensionality of the data set.

Feature selection procedures fall into three categories [58]: filter, wrapper and embedded. Filter methods act as a preprocessing step, implemented before the chosen model, and result in the removal of unwanted features to reduce the dimensionality of data input to the model. Filtering procedures include statistical tests for correlation between feature and outcome, and retaining a pre-defined set of features thought to be informative by mechanistic insight.

Wrapper methods consist of the identification of a series of feature subsets, and repeated model fitting using these feature subsets. Model fit is assessed using a chosen score, such as Akaike information criterion (AIC), Bayesian information criterion (BIC) or correlation between predictions and true values, and the suitability of the feature set is recorded. These methods may use any model for the underlying fitting. Popular wrapper methods include stepwise feature selection, genetic algorithms and simulated annealing. Wrapper methods are often found to be computationally costly, due to the repeated application of the model.

Finally, embedded methods are incorporated into the model itself. These methods include l_1-regularization techniques such as elastic net penalisation for generalised linear models [45], decision trees such as Random Forest [20] and the automatic relevance determination kernel for Gaussian processes [127]. Embedded methods require that the whole data set is passed to the model, and hence dimensionality reduction is not usually possible. However, the embedded feature selection typically increases the model complexity, which brings with it challenges for computation and inference.

One simple form of feature selection is the identification of a set of interesting features before data interrogation, using prior knowledge of the real-life context. For medical data, this could be the underlying disease mechanisms or known responses to treatment. In Chapter 4, two gene lists are specified using biological knowledge from literature. The first was composed of genes identified as mechanistically relevant to chemoresistance in ovarian cancer. The second consists of genes found to be predictive by two or more studies identified during a systematic review of the literature concerning the prediction of resistance to chemotherapy in ovarian cancer using gene expression, shown in Chapter 2. These gene sets were used in Chapter 4 for feature selection of high dimensional gene expression data sets prior to model fitting.

For Gaussian processes, emphasis is often placed on the computational dependence on the number of samples. However, computational complexity is also typically dependent on the number of dimensions, though the covariance function. It

is therefore important that the number of dimensions be reduced, both to alleviate the curse of dimensionality and improve run time.

Feature selection for Gaussian processes is not a widely studied area and there are few general methods for feature selection with Gaussian processes. The automatic relevance determination (ARD) covariance function is a highly used approach to feature selection [127]. This covariance function involves a hyperparameter per feature, and by varying these hyperparameters the relationship between each feature and the response may be changed. However, due to the large number of hyperparameters this method may be prone to overfitting. Many alternative methods implement modifications of a spike-and-slab prior (notably described by George and McCulloch [49]) on the ARD covariance function [89, 135, 136], with the aim of setting features as constant. Other techniques tend to employ Gaussian processes as a component of a more complex feature selection procedure [122, 123, 26]. It is therefore of interest to investigate feature selection in the context of Gaussian processes.

In Chapter 5, two feature selection methods for Gaussian processes are investigated. The first method, an embedded method, takes the form of a modification of the ARD kernel. This method utilises prior knowledge of feature similarity and relevance to group features and reduces the resulting number of model hyperparameters. The second method acts as a wrapper to allow Bayesian model averaging using randomly selected feature subsets. The resulting models are used to produce ensemble predictions, greatly reducing the computational requirements compared to the full, all-features model.

1.8 Thesis chapters

The chapters of this thesis are as follows:

Chapter 2 details a published systematic review into the literature concerning the prediction of resistance to chemotherapy in ovarian cancer using gene expression measurements.

Chapter 3 introduces three Gaussian process models designed to model censored survival times as partially missing outcomes, using covariates to place Gaussian process priors on the space of all functions relating covariates to outcome. These models are implemented and investigated using synthetic and molecular data sets in Chapter 4. They are observed to have equal or superior predictive ability compared to commonly used alternative models.

Two feature selection procedures for Gaussian processes are introduced and

implemented in Chapter 5. The two methods are tested using synthetic and molecular data sets and again compared to suitable alternative models. These methods were observed to have equal predictive ability to the comparison models.

Chapter 6 describes an interactive analysis program developed to analyse data produced by DNA-expression-based cancer mutation testing, and produce reports for clinical use detailing the mutations observed. This program applies traditional q-PCR analysis techniques, and these are carried out in a stand-alone manner, allowing the analysis to be carried out reliably and without the requirement of knowledge of the process.

Chapter 2

Systematic Review: Prediction of resistance to chemotherapy in ovarian cancer

This chapter presents a systematic review that investigates the literature concerning the prediction of resistance to chemotherapy in ovarian cancer using gene expression measurements. Following data compilation, some analysis was then conducted to assess the reliability, applicability and conclusions of the included studies, both on a study-by-study basis and as meta-analysis.

A version of this work has been published as [92]:
Katherine L Lloyd, Ian A Cree, and Richard S Savage. Prediction of resistance to chemotherapy in ovarian cancer: a systematic review. *BMC Cancer*, 15(1): 117, 2015.

Ian Cree and Richard Savage conceived and planned the study. Literature searches and analysis were carried out by Katherine Lloyd.

2.1 Introduction

2.1.1 Systematic Review

When starting research into a new topic, it is vital to gain a comprehensive view of the existing literature. The benefits of this approach are two-fold. Firstly, this allows the breadth and thoroughness of existing research into the topic of interest to be seen. In this way, gaps in the literature and hence prospective research areas become clear. Secondly, a thorough investigation of the existing literature provides an opportunity

to assess the success of existing studies. It is important that study effectiveness, bias and limitations are acknowledged when compiling conclusions based on existing literature. In addition, knowledge of the strengths and shortcomings of previous work should also guide the defining of a research area.

Rather than a general literature review, here it was decided that a systematic review was appropriate, as this format allows a narrow question to be investigated thoroughly. Due to the availability of publication databases, comprehensive literature searches are also increasingly feasible and simple to perform.

When carrying out a systematic review, a basic set of steps are followed. This structured approach allows the results to be reproducible and hence provides support to conclusions drawn following meta-analysis. As listed by Khan et al. [80], the five steps are:

- Framing the question
- Identifying relevant publications
- Assessing study quality
- Summarising the evidence
- Interpreting the findings

Following the identification of a focused, definable research question, search terms for the databases of choice are developed. These terms are modified to provide a list of papers that is hoped to be complete and exhaustive, without becoming too general. Once the searches are carried out, the list of papers is then filtered, by reading and checking that the search criteria are fulfilled. For speed, this is often carried out twice, initially using abstracts and again using full articles. Any papers found not to fulfil the search criteria are rejected and discounted from further analysis. The gathered papers are then assessed in detail. Methodological aspects and results are compiled for analysis, both on a paper-by-paper basis and as meta-analysis, allowing conclusions to be drawn. Acknowledgement of possible bias and limitations of studies are also important here, as this will affect the confidence put in results and conclusions.

2.1.2 Research Question

Gene expression based tools for the prediction of patient prognosis after surgery or chemotherapy are currently available for some cancers. For example, MammaPrint® uses the expression of 70 genes to predict the likelihood of metastasis in breast cancer

[146]. Similarly, the Oncotype DX® assay uses the expression of a panel of 21 genes to predict recurrence after treatment of breast cancer [7]. The Oncotype DX assay is also available for colon [8] and prostate cancers [9]. The development of a similar tool for ovarian cancer could greatly improve patient prognosis and quality of life by guiding chemotherapy choices. The prediction of cancer prognosis using gene signatures is a popular research field, within which a wide variety of approaches have been considered.

For ovarian cancer in the UK, the standard of care for first-line chemotherapy treatment recommended by the National Institute for Health and Care Excellence (NICE) is 'paclitaxel in combination with a platinum-based compound or platinum-based therapy alone' [105]. This uniform approach ignores the complexity of ovarian cancer histologic types, particularly as there is evidence to suggest differences in response [55, 166].

Improvement in survival has been poor in ovarian cancer. Between 1971 and 2007 there was a 38% increase in relative 10-year survival in breast cancer, whereas the increase in ovarian cancer was 17% [2]. This difference in progress is likely to be due, at least in part, to the lack of tools with which to predict chemotherapy response in ovarian cancer.

The computational and statistical approaches employed in studies predicting chemotherapy response vary greatly. One popular method is Cox proportional hazards regression. This model assumes that the hazard of death is proportional to the exponential of a linear predictor formed of the explanatory variables. This model has the advantage that, unlike many other regression techniques, it can model right-censored data such as that found in medical studies where patients leave before the end of the study period [40]. This technique is therefore often used in medical contexts, sometimes to the exclusion of other more suitable techniques. Other popular modelling techniques include linear models, support vector machines, hierarchical clustering, principal components analysis and the formation of scoring algorithms.

Models of survival data may be predictive or prognostic. For prognostic models the emphasis is on the prediction of patient survival time with no reference to treatment. Predictive models, on the other hand, take into account treatment and predict survival time given a chosen therapeutic intervention [160]. This distinction is subtle and, clearly, prognostic models may be formulated as predictive models given no treatment. However, in practice the difference is contextually important. At diagnosis or following an intervention, prognostic models with an emphasis on non-clinical covariates may be used to predict patient survival and disease progression [167]. However, following tests and treatment, predictive models utilising test results

and clinical disease assessments may be of use to tailor interventions and aid treatment selection. In the context of patient response to chemotherapy, given the varying biological processes and mechanisms of action involved, predictive models are much more likely to be effective, predicting patient response given a chosen treatment.

The aim of this review is to investigate the literature surrounding the prediction of chemotherapy response in ovarian cancer using gene expression. It has been observed, for example by Gillet et al. [53], that gene signatures obtained from cancer cell lines are not always relevant to *in vivo* studies, and that cell lines are inaccurate models of chemosensitivity [31]. The search was therefore restricted to studies involving human tissue in order to ensure that the resulting gene signatures are applicable in a clinical setting. It was also specified that the study must involve patients who have undergone chemotherapy treatment, so that the effects of resistance may be investigated.

A PRISMA checklist was completed for this document and may be found in Appendix Figure A.2.

2.2 Methods

2.2.1 Search Methodology

The aim of this review is to investigate the literature on the prediction of chemoresistance in patients with ovarian cancer. Therefore, the six most important requirements identified were:

- Concerned with (specifically) ovarian cancer
- Patients were treated with chemotherapy
- Gene expression was measured for use in predictions
- Predictions are related to a measure of chemoresistance (e.g. response rates, progression-free survival)
- Measurements were taken on human tissue (not cell lines)
- The research aim is to develop a diagnostic tool or predict response

A PubMed search was carried out on 6th August 2014 to identify studies fulfilling the above requirements. The search terms may be found in Appendix Table A.1. This search resulted in 78 papers.

2.2.2 Filtering

During reading, each abstract or full text was assessed as to whether it was appropriate for the research question and those not found to be suitable were excluded. The search results were filtered twice, once based on abstracts and once based on full texts. An overview of the filtering process may be found in Figure 2.1. For the abstract-based filtering, papers were excluded if the six essential criteria were not all met, if the paper was a review article or if the paper was non-English language. This resulted in 48 papers remaining. For the full-text-based filtering, exclusion was due to not fulfilling the search criteria or papers that were not available. 42 papers were remaining after full-text-based filtering.

2.2.3 Data Extraction

Data was extracted from each paper using a pre-defined table created for the purpose, with a number of well-defined fields of information to be extracted. Extraction was carried out in duplicate with a wash-out period of 3 months to avoid bias. The wash-out period, a period of time between two extractions by the same investigator, is intended to allow the second round of extraction to be approached fresh, to mimic extraction by two independent investigators. Variables extracted were:

- Author
- Year
- Journal
- Number of samples
- Number of genes measured
- Study end-point
- Tissue source
- Percentage cancerous tissue
- Gene or protein expression measurement technique
- Sample histological types
- Sample histological stages
- Patient prior chemotherapy

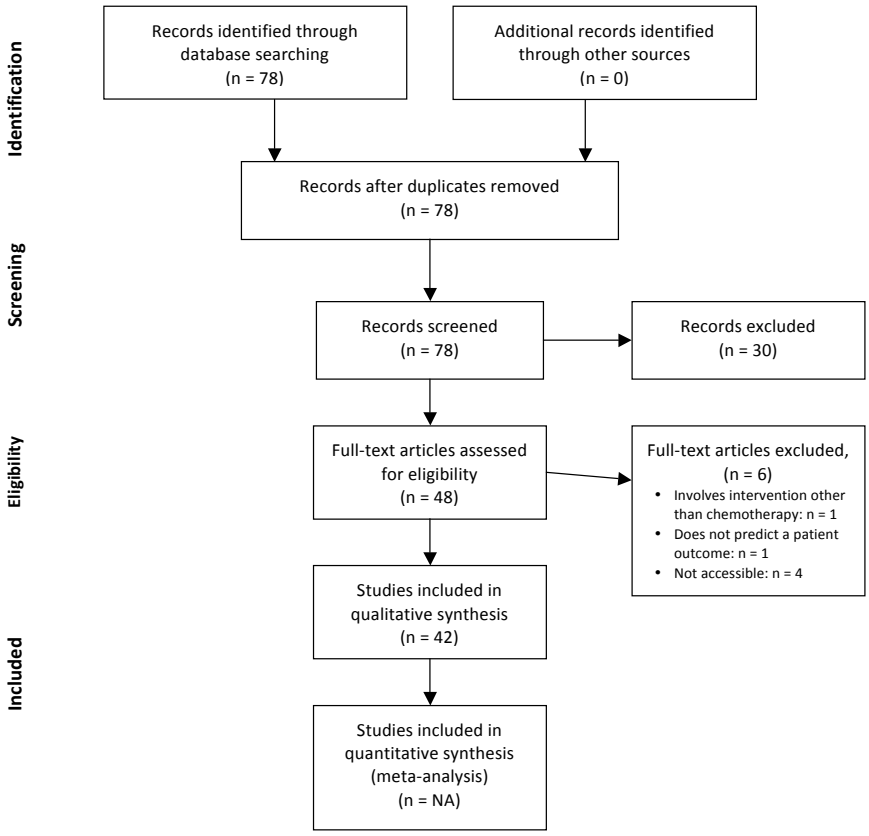


Figure 2.1: PRISMA search filtering flow diagram. The initial search results were filtered using titles and abstracts and, later, the full text to ensure the search criteria were fulfilled. Following filtering the number of papers included reduced from 78 to 42.

- Modelling techniques applied
- Whether the model accounts for heterogeneity in patient chemotherapy
- Whether the model was prognostic or predictive
- Whether the model was validated
- Model predictive ability including any metrics or statistics
- Genes found to be predictive

2.2.4 Bias Analysis

Bias in the studies selected for the systematic review was assessed according to QUADAS-2 [164], a tool for the quality assessment of diagnostic accuracy studies. For each paper, this tool assesses the risk of bias and the applicability of the study to the research question being investigated by the systematic review. There are four topics: patient selection, index test, reference standard, and flow and timing. Each topic is considered for risk of bias and applicability and assigned a rating of 'high risk', 'low risk' or 'unclear'. Patient selection is concerned with randomness, representativeness and exclusions. The index test is the test of interest to the review and the reference standard is the current test used for this purpose. Flow and timing considers the timing of the study, when tests were carried out and which patients received which tests. For each paper, results should be considered across the topics, with a paper only being considered low risk if all topics have achieved this status.

Levels of evidence were also assessed according to the CEBM 2011 Levels of Evidence [3]. This system ranks studies and clinical trials using several factors to assess the strength of their evidence. Different types of studies are considered to provide different strengths of evidence and are assigned levels of 1 to 5 accordingly, with 1 providing the strongest evidence. For example, systematic reviews and random controlled trials are considered to provide strong evidence, whereas case-control studies provide much less.

Results of these analyses may be found in Appendix Figure A.3. Briefly, the majority of studies were considered to be low risk, with six studies judged to have unclear risk for at least one domain and seven studies judged to be high risk for at least one domain. Thirty-six studies were judged to have evidence of level 2, with the remaining six having evidence of level 3. These levels of risk and evidence suggest that the majority of conclusions drawn from these studies are representative and applicable to the review question.

2.2.5 Gene Set Enrichment

Gene set enrichment analysis (GSEA) was applied to the gene sets reported by the studies selected for this review.

GSEA is a technique by which large sets of genes are assessed for their functional associations with known biological processes. This is often used to investigate links to disease phenotypes, for example. GSEA uses pre-existing relational databases of classifications to categorise genes, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. An enrichment score is calculated for each category in the database, showing to what extent it is represented in the provided set of genes, compared to the full population of genes included in the chosen database. For example, if genes associated with the cell cycle are relatively more common in the gene set than in all genes, this category will have a larger enrichment score and be considered over-represented. The hypergeometric test is used to assess whether a particular category has been observed statistically significantly more often than for a random sample drawn from the set of all genes. This test is applied for each category, and hence a p-value correction is required to adjust for multiple testing.

Analysis was performed using the R package *HTSanalyzeR* [162]. Where reported by studies included in the review, gene sets were extracted and combined according to the chemotherapy treatments applied to patients in each study. The two groups assessed were those studies where all patients were treated with platinum and taxane in combination, 'platinum and taxane', and those studies where patients were given treatments other than platinum and taxane, 'other treatments'. The 'other treatments' group includes those given platinum as a single agent. Any studies reporting treatments from both groups were excluded, as were studies that did not report the chemotherapy treatments used. Using functions from the *HTSanalyzeR* package, KEGG terms were identified for each gene and gene set collection analysis was carried out, which applies hypergeometric tests and gene set enrichment analysis. A p-value cut-off of 0.0001 was used. Enrichment maps were then plotted, using the 30 most significant KEGG terms. P-values were adjusted using the Benjamini-Hochberg correction [18]. On the enrichment maps, each node represents a set of genes related to the named KEGG term, and the size of the node is determined by the number of genes present related to that term. The edges link related gene sets, and their thickness represents a measure of similarity between the two sets of genes, according to a weighted Kolmogorov–Smirnov-like statistic. The nodes are coloured according to the p-value reported for the hypergeometric test for that term, and darker colour represents greater significance.

2.3 Results

Appendix Tables A.1, A.2, A.4, A.3 A.5 and A.6 detail some key information regarding the studies included in the review. Table A.1 contains the number of samples analysed, the number of genes considered for the model, and the resulting genes retained as the predictive gene signature. Table A.2 provides information about the tissue used for gene expression measurements and whether the studies assessed the percent neoplastic tissue before measurement, and Table A.4 details the gene expression measurement techniques used. Table A.3 contains the reported histological types and stages of the samples processed by each study. Table A.5 provides information on chemotherapy treatments undergone by patients, whether the model was prognostic or predictive, and whether the model was validated using either an independent set of samples or cross validation. Table A.6 lists the outcome to be predicted, the modelling techniques applied, and the predictive ability of the resulting model.

2.3.1 Tissue Source

For studies involving RNA extraction the tissue source is an important consideration, as RNA degradation and fragmentation could affect the results of techniques involving amplification. This is a notable issue in formalin fixed paraffin embedded (FFPE) tissue, due to the cross-linking of genetic material and proteins [102]. Of the 42 papers included in this review, the majority used fresh-frozen biopsy tissue. The numbers of each tissue source may be found in Table 2.1, and the tissue source used by individual papers may be found in Appendix Table A.2. Nine papers did not use an RNA source directly as secondary data was used. Data sources were mostly other studies or data repositories, such as the TCGA dataset. Two studies did not specify the source tissue though extraction and expression measurement methods were detailed.

The majority of papers in this review used fresh-frozen tissue. This choice was likely made to minimise RNA degradation and hence improve measurement accuracy. Due to the risk of RNA degradation because of long storage times and the fixing process applied to FFPE tissue, it is often expected that FFPE tissue will be irreversibly cross-linked and fragmented. However, following investigation into RNA integrity when extracted from paired FFPE and fresh-frozen tissue, Rentoft et al. [129] found that for most samples up- and down-regulation of four genes was found to be the same whether measured in FFPE or fresh-frozen tissue. They concluded that, if samples were screened to ensure RNA quality, FFPE material can successfully

Table 2.1: Numbers of studies using various mRNA sources.

mRNA Source	Number of studies
FFPE tissue	12
Fresh-frozen tissue	22
Fresh-frozen effusion	2
Fresh tissue	1
Blood	1
Not used	9
Not specified	2

provide RNA for gene expression measurement.

The use of fresh-frozen tissue in a research setting is not unusual, as can be seen from the fact that this tissue type was most popular in this review. However, for translational research expected to lead to a clinical test, this is not as reasonable. FFPE tissue is much more readily available, due to simpler acquisition and storage, and tissue is already taken for histological analysis. Therefore a model capable of using data obtained from FFPE tissue is much more likely to be applicable in a clinical setting.

Another important consideration is the proportion of neoplastic cells in the sample. For each paper the reported proportion may be seen in Appendix Table A.2. Of the 42 papers, 14 reported that the proportion of cancerous cells was measured. This was usually done using hematoxylin and eosin stained histologic slides. It is important for the gene expression measurement that the tissue used contains a high proportion of neoplastic cells, and hence it is important that this pre-analytical variable is controlled. Of the studies in this review, those reporting the percentage cancerous cells were evenly distributed between FFPE and fresh-frozen tissues.

2.3.2 Gene or Protein Expression Quantification

Of the studies highlighted by this review, there were four main techniques applied for gene or protein expression measurement: Probe-target hybridization microarrays, quantitative PCR, reverse transcription end-point-PCR, and immunohistochemical staining. Of these methods only immunohistochemistry measures protein expression, via classification of the level of staining, and the other methods quantify gene expression via measurement of mRNA copy number.

Methods involving probe-target hybridization are available commercially, and 19 of the 42 studies utilised these. For example the Affymetrix® Human U133A 2.0

GeneChip and the Agilent® Whole Human Genome Oligo Microarray were both used by multiple studies. Additionally, 7 studies used custom-made probe-target hybridization arrays. Probe-target hybridisation arrays generally measure thousands of genes and hence can provide a wealth data per sample. TaqMan® microfluidic arrays or quantitative-PCR were used by 16 studies. These techniques are typically used for smaller panels of genes. The TaqMan® arrays, for example, may contain up to 384 genes per array. These methods are more targeted and hence the price per sample is usually lower.

Immunohistochemistry is a more labour-intensive technique, requiring staining for each gene considered, and hence was mostly only used by studies using small numbers of genes. This technique, which is semi-quantitative due to the scoring systems employed, also suffers from a lack of standardisation of procedures. Of the 11 papers using this technique, the maximum number of genes analysed was seven, and the mean number of genes assessed was 2.8. Although these studies provide useful information regarding the correlation of particular genes with outcome, the small numbers of genes is likely to result in an incomplete gene signature and low predictive power.

Several of the papers utilising quantifiable techniques used an alternative method or replicates to obtain a measure of the assay variability. Five papers involving commercial or custom microarrays also used reverse transcription PCR (RT-PCR) to measure the expression of a small number of genes for comparison and one study used samples run in duplicate to calculate the coefficient of variation. Of the studies using TaqMan microfluidic arrays, two used samples run in duplicate to obtain the coefficient of variation. However, even fewer papers reported a metric representing the level of variability found. Two studies reported a coefficient of variation; Glaysher et al. [54] reported CoV = 2% = 0.02 for TaqMan arrays and Hartmann et al. [63] reported CoV = 0.2 for their custom microarray.

Another two reported Spearman's ρ or Pearson's r coefficients of correlation between microarray and RT-PCR results. Both coefficients represent the correlation between two variables, here two alternative mRNA measurement techniques. These coefficients may take values between -1 and 1, with 1 representing perfect positive correlation, 0 being no correlation, and -1 being perfect negative correlation. However, they differ in that Pearson's r assesses linear correlation, whereas Spearman's ρ measures monotonic correlation. Yoshihara et al. [170] gave Pearson's r values ranging from 0.5 to 0.8, and Crijns et al. [32] gave Spearman's ρ values between -0.6 and -0.9. Comparing these values to the possible ranges, both studies appear to have good correlation between techniques.

2.3.3 Histology

Appendix Table A.3 details the histology (types and stages) of the patient samples used by each study. As may be seen, the majority of studies were heterogeneous with respect to the types of cancer included. However, 23 of the 42 studies used at least 80% serous samples, suggesting that the majority of information contributed to the gene signatures of these studies is related to the mechanisms and pathways in serous cancer. It is important to identify the histologies of patient samples: although treatment is currently the same across types, response to chemotherapy has been found to vary [166, 150, 68]. It therefore may be advisable for future studies to include histological information when developing models predicting chemotherapy response.

2.3.4 Chemotherapy

Appendix Table A.6 lists the chemotherapy treatments undergone by patients in each study. The 10 papers labelled NS did not specify the regimen applied, though the patients did have chemotherapy. These cohorts cannot therefore be assumed to be homogeneous with respect to patient chemotherapy treatment. All studies that specified the chemotherapy regimen undergone by patients noted at least one platinum-based treatment. Of these, 24 included patients treated with a platinum-taxane combination and 10 with a cyclophosphamide-platinum combination. It is important to note that 19 of the 42 papers stated the population was heterogeneous with regards to chemotherapy treatments and, of those that did, only 8 included patient treatment history as a feature of the study. The aims of the majority of the studies were to identify genes of which the expression may be used to predict survival time, or prognosis. As already noted, the presence of resistance to the chemotherapy agent administered will dramatically affect the survival of a patient. It is therefore reasonable to expect the gene signatures identified to include genes responsible for chemoresistance, which will depend on the mechanism of action of the drug. Using a heterogeneous cohort in terms of chemotherapy treatment may then be causing problems with the identification of a minimal predictive gene set.

2.3.5 End-point to be Predicted

As may be expected, there was variation between the end-point chosen by studies for prediction. Popular end-points include overall survival, progression-free survival and response to chemotherapy. The endpoints considered by each study may be found in Appendix Table A.6. Of these some are clinical endpoints, such as overall

survival, others use non-clinical endpoints, such as response to chemotherapy, many of which are considered to be surrogates for overall survival. For cancer studies, overall survival is considered to be the most reliable and is the variable that is of most interest when considering the effect of an intervention.

2.3.6 Model Development

Within this review, many different modelling techniques were used to identify an explanatory gene signature to predict patient outcome. The most popular was Cox proportional hazards regression, which was applied by 17 studies. This was closely followed by hierarchical clustering, which was used by 11 studies. All other methods were used by 8 or fewer studies. In total 24 different types of modelling techniques were applied, ranging from statistical tests such as Student's t test and the Mann-Whitney U test, to logistic regression and ridge regression. Table 2.2 lists the modelling techniques identified and the number of studies that employed them. It is of interest that most of the techniques applied are forms of classification. These methods result in samples being assigned to groups, such as 'good prognosis' and 'poor prognosis'. Whilst this may be useful in some settings, for a clinically-applicable tool a regression technique may be more appropriate as it will provide a value, such as a likelihood of relapse, rather than simply a class. Techniques in Table 2.2 capable of a numeric prediction include logistic and linear regression, Cox proportional hazards regression, and ridge regression.

Jointly with the modelling methods identified above, 23 of the 42 studies implemented Kaplan-Meier curves to visualise the survival of the patient classes identified by the models. This enables the difference in survival between classes, for example 'good prognosis' and 'poor prognosis', to be seen and assessed. The application of a log-rank test assesses the separation of the curves and identifies whether there is a statistically significant difference in survival distribution between the classes. It should be noted that, although this gives an idea of separation of classes achieved by the model, the model results must still be compared with known outcomes to check positive and negative predictive power. This step was missing in several papers, such as Gillet et al. [51], where the p-value returned by the log-rank test is given as the measure of model success.

It is important to highlight the difference between prognostic and predictive models. A prognostic model is one capable of predicting prognosis, such as survival time, using patient information and biomarkers and does not vary between different treatment options. In contrast, a predictive model is one able to predict the effect of a treatment on patient prognosis [121, 153]. It is therefore clear that, although

Table 2.2: Key Modelling techniques applied by studies in the review.

Technique	Number of papers
Cox proportional hazards regression	17
Hierarchical clustering	11
Principal components analysis	8
Student's t test	7
Scoring algorithm	6
Support Vector Machines	5
Correlation coefficients	5
Mann-Whitney U test	5
χ^2 test	5
ROC analysis	5
Class prediction	4
Logistic regression	3
Linear regression	3
AIC gene selection	2
Concordance index	1
Pathway interaction networks	1
ANOVA	1
Expression threshold identified	1
Gene set enrichment analysis	1
Linear discriminant analysis	1
ISIS bipartitioning	1
Gaussian mixture modelling	1
Significance analysis of microarrays	1
Ridge regression	1

prognostic models may be useful for research purposes and when only one treatment option is available (such as the standard platinum-taxane combination), predictive models have a much greater part to play in stratified medicine where the aim is to identify the most appropriate treatment on a patient-by-patient basis. In order for a model to be predictive, the effects of multiple treatments must be considered and the response compared with the biomarker status. Classification of the studies as prognostic or predictive may be seen in Appendix Table A.5. Of the papers identified by this review, only a minority considered the effects of chemotherapy treatment on the predicted outcome and hence could be considered predictive. Glaysher et al. [54] and Vogt et al. [161] produced separate models for various treatments, allowing the effects of different drugs and combinations to be compared. Both studies applied drugs *in vitro* to cultured tissue to measure response to chemotherapy. This was combined with gene expression measurements to form the model training data set. In this way the same patient samples may be used to create a set of models predicting response to a variety of drugs. These models are therefore predictive rather than prognostic. Alternatively, models may be trained on sets of patients split by treatments undergone, which would lead to treatment-specific models predicting response to the particular drug. This method was used by Jeong et al. [76], Ferriss et al. [42], Williams et al. [165] and Matsumura et al. [98]. Additionally, the use of a model variable specifying patient treatment history could allow these models to be combined onto one using a single training set of all patients. The model may then be passed a variable specifying the drug of interest for resistance prediction. A simple version of this method was implemented by Crijns et al. [32], who included a feature for whether a patient was treated with paclitaxel. It is clear that the integration of patient chemotherapy treatment into these models is underused, and it is likely to be beneficial for this to be incorporated into future research.

2.3.7 Genes Identified

Of the 42 papers in this review, 32 provided full or partial lists of the genes identified by their models. Of the remainder, it was common that the gene sets were large or that the genes were not explicitly identified by the model, as is the case with modelling techniques such as principal components analysis.

In total across the papers, 1298 unique genes were selected by models and of these 93.53% were found by only one paper. The most commonly chosen gene was selected by only four papers. Table 2.3 shows the numbers and percentages of genes chosen by one to four papers.

A list of the genes identified by the papers in the review may be found in

Table 2.3: Numbers and percentages of genes featured in the gene sets of various numbers of papers.

Number of papers identifying a gene	Number of genes	Percent of genes
1	1214	93.53%
2	78	6.01%
3	5	0.385%
4	1	0.08%

Appendix Table A.7.

It is clear that the gene sets selected by the studies are very different and there is very little overlap. The genes chosen by two or more studies may be seen in Appendix Table A.8. Many of these genes are known to have links to cancer, which may suggest that these genes are therefore implicated in ovarian cancer. It is possible that, although the genes selected varied, they in fact represent similar mechanisms. This could occur if there are large sets of highly covariate genes representing particular cellular processes and the genes in the signatures were simply random selections from these gene sets. The same gene being selected by multiple papers would then be unlikely, although the same information contribution would be made. It may then be more informative to assess and compare the mechanisms controlled by the genes chosen as part of the models.

2.3.8 Gene Set Enrichment

The gene sets reported by the studies identified in this review were assessed to identify whether certain biological pathways and mechanisms featured more prominently according to the genes selected. As detailed in Methods section 2.2.5, the genes selected by the studies are compared to the set of all genes and their associated pathways in the KEGG database. Pathways identified more commonly than in the full population are considered enriched.

In order to explore the differences between chemotherapy treatments, studies were grouped by chemotherapy treatments recieved by the patients. The two groups identified were 'platinum and taxane', and 'other treatments' (such as platinum, cyclophosphamide and combinations). Studies that did not specify the chemotherapy treatments used were excluded. Studies falling into the 'platinum and taxane' group were Han et al. [61], Kang et al. [78], Gillet et al. [52], Skirnisdottir and Seidal [145], Schlumbrecht et al. [139], Yoshihara et al. [170], Denkert et al. [37], Hartmann

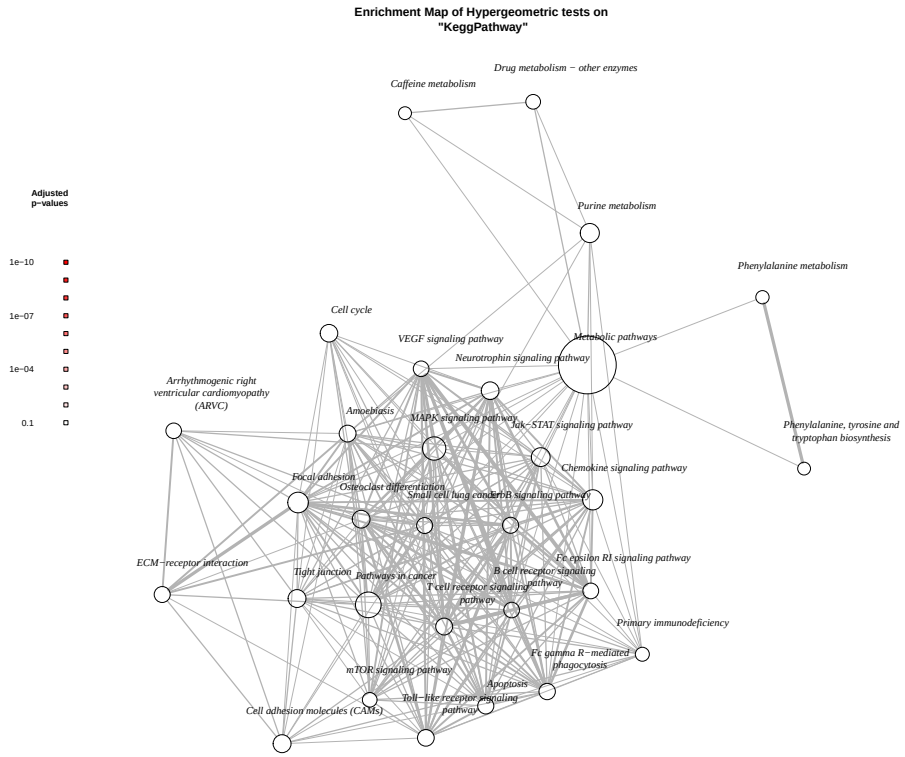
et al. [63], Iba et al. [72], and Kamazawa et al. [77]. Studies falling into the 'other treatments' group were Obermayr et al. [120], Obermayr et al. [120], Yan et al. [169], Netinatsunthorn et al. [115], and Helleman et al. [66]. The results of the gene set enrichment using the KEGG system may be seen in Figures 2.2a and 2.2b. From the plots, it may be seen that both groups identify several cancer-related pathways relevant to the drug mechanisms of action.

It is informative to consider the KEGG terms in the context of the mechanisms of action of the chemotherapy drugs applied. Both groups contain patients treated with platinum single agents or platinum-containing combinations. It should therefore be expected that processes associated with the mechanism of action of platinum will be enriched. Once activated, the platinum binds to DNA and results in the formation of monoadducts, intra-strand crosslinking, inter-strand crosslinking and protein crosslinking. This DNA structure change affects the ability of the DNA to be unwound and replicated, resulting in the triggering of the G2/M DNA damage checkpoint and cell cycle arrest. The affected cell will attempt DNA repair and, if unsuccessful, undergo apoptosis [152]. Expected KEGG terms therefore include those relating to apoptosis and DNA damage.

From Figure 2.2a, KEGG pathways highlighted for this group of studies include ten cancer-specific terms and six cancer-related terms. Here italics denote a KEGG term. The *ErbB signalling pathway* has been found to influence proliferation, migration, differentiation and apoptosis in cancer [27] and overexpression of ERBB1 and ERBB2 have been implicated in head and neck and breast cancers. The *neurotrophin signalling pathway* is known to trigger MAPK and PI3K signalling, affecting differentiation, proliferation and development, and survival, growth, motility and angiogenesis respectively [151]. Altered expression of genes in this pathway has been found to correlate with poorer survival in colon, breast, lung and prostate cancers. Changes in expression of genes relating to *focal adhesion*, which is responsible for attachment of cells to the extracellular matrix, have been implicated in cancer migration, invasion, survival and growth [100]. The *TGF-beta signalling pathway* also regulates many cellular processes, including proliferation, cellular adhesion and motility, coregulation of telomerase function, regulation of apoptosis, angiogenesis, immunosuppression and DNA repair [41]. The *p53 signalling pathway* has many varied links to cancer. This pathway many be triggered by various stress signals and can result in several responses, including cell cycle arrest, apoptosis, the inhibition of angiogenesis and metastasis, and DNA repair [87]. Finally, *nucleotide excision repair* is known to promote cancer development when both up and down regulated. Down-regulation is thought to increase susceptibility to mutation formation and

(a) Gene set enrichment networks for studies assessing ovarian cancer patients treated with platinum and taxane.

Figure 2.2: Network maps of the 30 most enriched KEGG pathways. Node marker size signifies the number of genes in this category, and the thickness of edges indicate the Jaccard similarity coefficient between categories. Node markers are coloured according to adjusted p-value as reported by the hypergeometric test, where darker red denotes more highly significant.



(b) Gene set enrichment networks for studies assessing ovarian cancer patients treated with treatments other than platinum and taxane.

Figure 2.2: Network maps of the 30 most enriched KEGG pathways. Node marker size signifies the number of genes in this category, and the thickness of edges indicate the Jaccard similarity coefficient between categories. Node markers are coloured according to adjusted p-value as reported by the hypergeometric test, where darker red denotes more highly significant.

hence the formation of cancer [119], whereas up-regulation has been found to correlate with resistance to platinum as the DNA damage caused by the chemotherapy agent is repaired [96].

The first group of studies considered patients treated with taxanes in addition to platinum. Taxanes act by stabilising tubulin, preventing the microtubule structure formation required for mitosis. This results in cell cycle arrest at the G2/M DNA damage checkpoint and apoptosis. Mechanisms for taxane resistance are, however, not well understood. Two suggested mechanisms include the increased expression of multidrug transporters, and changes in the expression of the β-tubulin isoforms [47]. Neither of these mechanisms seem to be enriched in the platinum and taxol group. In addition to the single-agent effects of platinum and taxanes, there is an additional synergistic effect [75]. However, this effect is also not well studied and hence the mechanisms by which this occurs are not clear.

The second group, as seen in Figure 2.2b, was composed of studies applying chemotherapy treatments other than platinum and taxanes. This group is heterogeneous with respect to chemotherapy treatment, and mainly consists of studies reporting treatment as 'platinum-based'. The other drug explicitly mentioned by studies in this group is cyclophosphamide. This drug is an alkylating agent and acts to form adducts in DNA [59]. This DNA damage triggers the G2/M DNA damage checkpoint, resulting in DNA repair or apoptosis. This suggests that the same DNA repair mechanisms related to platinum treatment are also relevant to cyclophosphamide. For this group, the KEGG pathway analysis shows that the gene set is enriched with 14 pathways related to cancer, in addition to two general cancer-related terms. The *mTOR signalling pathway* is downstream to the PI3K/AKT pathway and regulates growth, proliferation and survival [124]. The *MAPK signalling pathway* controls the cell cycle, and has been found to contribute to the control of proliferation, differentiation, apoptosis, migration and inflammation in cancer [38]. The *chemokine signalling pathway* has been found to regulate growth, survival and migration in addition to its role in inflammation [67]. Angiogenesis and vasculogenesis are known to be regulated by the *VEGF signalling pathway* [84], which is already the target of treatments such as bevacizumab. *Purine metabolism* is required for the production and recycling of adenine and guanine, and hence is required for DNA replication. This process is the target of chemotherapies such as methotrexate. The term *drug metabolism – other enzymes* is partially cancer related; this term refers to five drugs: azathioprine, 6-mercaptopurine, irinotecan, fluorouracil and isoniazid. Of these, two are chemotherapy treatments; irinotecan is a topoisomerase-I inhibitor and fluorouracil acts as a purine analogue. Also featuring in Figure 2.2b are *apoptosis*,

ErbB signalling pathway, *focal adhesion*, *neurotrophin signalling pathway*, *B cell receptor signalling pathway* and *Jak-STAT signalling pathway*, all of which are known to be related to cancer.

Overall, the gene sets appear to be enriched for cancer-related resistance mechanisms [30], though very few of the terms were considered to be statistically significantly enriched. However, when combined there is little evidence from this analysis to suggest that the signatures are capturing chemotherapy-specific mechanisms in addition to more general survival pathways. The DNA repair terms may suggest a response to platinum-based treatment, though the down-regulation of these mechanisms is also related to cancer development and resistance in general [19]. It is likely that, due to the varying reliability suggested by the bias analysis and the reported model development techniques, the signal-to-noise ratio of informative genes is low when the gene signatures are combined, preventing the identification of processes of interest.

2.3.9 Model Predictive Ability

Sensitivity and Specificity

The comparison of the success of the various models is difficult, particularly due to the fact that many papers report different metrics as measures of model accuracy. Many of these are also incomplete, not providing enough information to fully describe the model. Ideally, models should be applied to an independent set of samples with known outcomes and performance measures on this data set reported. For classification models an informative set of measures would be positive predictive value (PPV), negative predictive value (NPV), specificity and sensitivity:

$$\text{Sensitivity} = \frac{n_{\text{true positive}}}{n_{\text{true positive}} + n_{\text{false negative}}}$$

$$\text{Specificity} = \frac{n_{\text{true negative}}}{n_{\text{true negative}} + n_{\text{false positive}}}$$

$$\text{PPV} = \frac{n_{\text{true positive}}}{n_{\text{true positive}} + n_{\text{false positive}}}$$

$$\text{NPV} = \frac{n_{\text{true negative}}}{n_{\text{true negative}} + n_{\text{false negative}}}$$

where $n_{\text{true positive}}$ is the number of true positive predictions, $n_{\text{false positive}}$ is the number of false positive predictions, $n_{\text{true negative}}$ is the number of true negative predictions and $n_{\text{false negative}}$ is the number of false negative predictions.

Together these provide information on true positive and negative rates as well

as false positive and false negative rates, all of which are important when assessing the performance of a model.

Using the sensitivity and specificity the positive and negative likelihood ratios may be calculated and, using the prevalence of the condition in the test population, the probability of a patient having the condition based on the test results may be found, as in the equations below. The likelihood ratios are the ratio of the probabilities of a patient testing positive (or negative), given that they have the disease or not.

$$\mathrm{LR}_{+\mathrm{ve}} = \frac{P(\mathrm{Test}+|\mathrm{Condition}+)}{P(\mathrm{Test}+|\mathrm{Condition}-)} = \frac{\text{sensitivity}}{1-\text{specificity}}$$

$$\mathrm{LR}_{-\mathrm{ve}} = \frac{P(\mathrm{Test}-|\mathrm{Condition}+)}{P(\mathrm{Test}-|\mathrm{Condition}-)} = \frac{1-\text{sensitivity}}{\text{specificity}}$$

$$P(\mathrm{Condition}+|\mathrm{Test}+) = \frac{\frac{P(\mathrm{Condition}+)}{1-P(\mathrm{Condition}+)} \cdot \mathrm{LR}_{+\mathrm{ve}}}{\frac{P(\mathrm{Condition}+)}{1-P(\ \mathrm{Condition}+)} \cdot \mathrm{LR}_{+\mathrm{ve}} + 1}$$

$$P(\mathrm{Condition}+|\mathrm{Test}-) = \frac{\frac{P(\mathrm{Condition}-)}{1-P(\mathrm{Condition}-)} \cdot \mathrm{LR}_{-\mathrm{ve}}}{\frac{P(\mathrm{Condition}-)}{1-P(\mathrm{Condition}-)} \cdot \mathrm{LR}_{-\mathrm{ve}} + 1}$$

These post-test probabilities are much easier to interpret and incorporate the prevalence of the condition. It should be noted that in order for the test to be applied in a clinical situation the pre-test probabilities used, $P(\mathrm{Condition}+)$ and $P(\mathrm{Condition}-)$, should be correct for the population of patients to whom the test will be applied. Here the sample prevalence (the proportion of subjects having the condition) from each study was used for convenience. However, it would be informative to recalculate $P(\mathrm{Condition}+|\mathrm{Test}+)$ and $P(\mathrm{Condition}+|\mathrm{Test}-)$ for the general population of ovarian cancer patients, as this would provide a better comparison between models.

Table 2.4 details the post-test probabilities of patients having a condition based on a positive or negative test result from the models developed by studies in this review. The papers appearing here are those that supplied sensitivity and specificity and the numbers of patients with and with without the condition, or alternative information allowing these to be calculated such as numbers of true and false positives and negatives.

From the table it may be seen that there is a great variety between the success of the models. For example, Kamazawa et al. [77] and Hartmann et al. [63] both achieved $P(\mathrm{Condition}+|\mathrm{Test}+) = 0.95$ on their respective samples of the population. This means that if a patient tests positive, there is a 95% probability that they

Table 2.4: Prediction metrics for studies reporting sensitivity and specificity.

Study	Prediction	Sensitivity	Specificity	LR_{+ve}	LR_{-ve}	$P(C+)$	$P(C-)$	$P(C+\mid T+)$	$P(C+\mid T-)$
Li et al. [88]	Chemoresistance	0.96*	0.23*	1.24	0.18	$\frac{22}{44}$	$\frac{22}{44}$	0.55	0.15
Obermayr et al. [120]	RFS	0.22*	0.85*	1.47	0.92	$\frac{46}{216}$	$\frac{170}{216}$	0.28	0.77
Ferriss et al. [42]	Chemoresponse	0.94*	0.29*	1.33	0.20	$\frac{85}{119}$	$\frac{34}{119}$	0.77	0.07
Sabatier et al. [134]	Prognosis	0.62*	0.62*	1.64	0.62	$\frac{194}{366}$	$\frac{172}{366}$	0.65	0.35
Yoshihara et al. [170]	PFS	0.64*	0.69*	2.06	0.52	$\frac{45}{87}$	$\frac{39}{87}$	0.69	0.30
Williams et al. [165]	Prognosis	0.77*	0.56*	1.75	0.41	$\frac{97}{143}$	$\frac{46}{143}$	0.79	0.16
Gevaert et al. [50]	Chemoresistance	0.67*	0.40*	1.12	0.82	$\frac{15}{45}$	$\frac{30}{45}$	0.36	0.62
Helleman et al. [66]	Chemoresistance	0.89*	0.56*	2.02	0.20	$\frac{9}{72}$	$\frac{63}{72}$	0.22	0.58
De Smet et al. [35]	Chemoresistance	0.71†	0.83†	4.29	0.34	$\frac{6}{13}$	$\frac{7}{13}$	0.79	0.29
Raspollini et al. [128]	Prognosis	0.79†	0.46†	1.45	0.47	$\frac{28}{52}$	$\frac{24}{52}$	0.63	0.29
Hartmann et al. [63]	Prognosis	0.86*	0.86*	6.14	0.16	$\frac{21}{28}$	$\frac{7}{28}$	0.95	0.05
Selvanayagam et al. [142]	Chemoresistance	1.00†	1.00†	∞	0.00	$\frac{4}{8}$	$\frac{4}{8}$	1.00	0.00
Kamazawa et al. [77]	Chemoresponse	1.00*	0.83†	6.00	0.00	$\frac{21}{27}$	$\frac{5}{27}$	0.95	0.00

* Value stated in reference,
† Value calculated,
C: condition presence,
T: test result,
RFS: Relapse Free Survival,
PFS: Progression Free Survival

are positive for the condition in question, which in these cases are 'responding to chemotherapy' and 'poor prognosis' respectively. In contrast, Obermayr et al. [120], Helleman et al. [66] and Gevaert et al. [50] only achieved $P(\text{Condition}+|\text{Test}+)$ of between 0.20 and 0.40. These results suggest that the tests are not able to predict the outcome of a patient any better than a random choice, and in the case of tests in the region of 0.20 it is likely that most patients are simply assigned to the same class.

The ability of tests not to commit type II errors and give false negatives is also important. Ferriss et al. [42] and Hartmann et al. [63] both achieved well in this regard, with $P(\text{Condition}+|\text{Test}-) = 0.07$ and $P(\text{Condition}+|\text{Test}-) = 0.05$ respectively. Several studies, by contrast, had very poor probabilities of false negatives; Obermayr et al. [120], Helleman et al. [66] and Gevaert et al. [50] all have $P(\text{Condition}+|\text{Test}-) > 0.5$, which suggests that these models give a false negative more often than a random assignment.

Kamazawa et al. [77] and Selvanayagam et al. [142] both achieved extremely impressive prediction abilities, as may be seen by the very large $P(\text{Condition}+|\text{Test}+)$ and very small $P(\text{Condition}+|\text{Test}-)$ values. However, these studies exemplify why care must be taken in assessing the predictive ability of models. Both studies calculated sensitivity and specificity based on only training set results and hence there is no way to judge the generalisability of the models. There is a tendency for models to perform better on the training set than any following independent data set to which it is subsequently applied. Secondly, the training set used by Selvanayagam et al. [142] is extremely small at eight patients and has a $50:50$ ratio of chemoresistant to chemosensitive patients. This sample is not representative of the population and hence the values of $P(\text{Condition}+|\text{Test}+)$ and $P(\text{Condition}+|\text{Test}-)$ will be skewed by unrepresentative $P(\text{Condition}+)$ and $P(\text{Condition}-)$.

Overall, the most successful model of this group is that by Hartmann et al. [63] as it makes predictions with good reliability and has been validated on an independent data set. The least successful models were Obermayr et al. [120], Helleman et al. [66] and Gevaert et al. [50]. These studies suffered from low ability to identify true positives and high probability of false positives, resulting in poor predictive ability.

Hazard Ratios

It is common for studies of survival to quote hazard ratios, comparing the results of clusters identified by classification models or relative-risk models such as Cox proportional hazards regression. These ratios represent the ratio of the probability of an event occurring to a patient in either of the two groups. The event is often death,

but could also be recurrence for example. The studies listed in Table 2.5 supplied hazard ratios as measures of predictive ability. The hazard ratios vary from 0.23 to 4.6 with the majority around 2 to 3. A hazard ratio that is not equal to 1 suggests that the variable has predictive ability, and a ratio of 4, for example, suggests that a member of the high-risk group is 4 times as likely to die within the study period than a member of the low-risk group. The study with the highest hazard ratio is Spentzos et al. [147], with HR = 4.6. This is closely followed by Raspollini et al. [128] with HR = 0.23 and Skirnisdottir and Seidal [145] with HR = 4.12. The confidence intervals on the hazard ratios of all the studies are large and, with the exception of Spentzos et al. [147], at the lowest edge the hazard ratio is very close to 1. This suggests that, although all these hazard ratios were found to be significant, some were close to not reaching the arbitrary $\alpha = 5\%$ significance level. Most notable are Roque et al. [131], Schlumbrecht et al. [139], and Denkert et al. [37]. These models would need further investigation to determine their predictive ability. Of the papers in this group, Spentzos et al. [147] appears to have the best predictive ability when classifying patients into two clusters with significantly different survival times.

Linear Regression

Two papers reported the success of a model assessed using linear regression: Glaysher et al. [54] and Kang et al. [78]. These studies plotted the predicted values or model score against the measured values and applied linear regression to obtain a line of best fit. The R^2 or R^2_{adj} of this line is then calculated to assess the discrimination of the model and represents the proportion of the variance in the dependant variable that is explained by the independent variable. R^2_{adj} also accounts for the number of variables and the sample size. Glaysher et al. [54] achieved $R^2 = 0.901$ ($R^2_{\text{adj}} = 0.836$) for a model predicting resistance to cisplatin using gene expression via cross-validation and Kang et al. [78] achieved $R^2 = 0.84$ for a model predicting recurrence-free survival in the data set on which it was derived. These values suggest a good level of predictive ability, both in terms of calibration and discrimination, with the model by Glaysher et al. [54] achieving the better predictions.

Cox Proportional Hazards Models

When studies identified by this review applied the Cox proportional hazards model to predict patient outcome, it was common for the main analysis of the model to be assessing whether the gene signature was found to be significant and whether the signature was an independent predictor. However, the application of this model to

Table 2.5: Prediction metrics for studies reporting hazard ratios.

Study	Prediction	Classes	HR	95% CI	Median survival	P-value
Jeong et al. [76]	OS	YA subgroup vs. YI subgroup	0.5	0.31–0.82		0.005
Roque et al. [131]	OS	High vs. low TUBB3 staining	3.66	1.11–12.05	707 days vs. not reached	0.03
Kang et al. [78]	OS	High vs. low score	0.33	0.13–0.86	1.8 years vs. 2.9 years	< 0.001
Skirnisdottir and Seidal [145]	Recurrence	p53 -ve vs. +ve	4.12	1.41–12.03		0.009
Schlumbrecht et al. [139]	RFS	EIG121 high vs. low	1.13	1.02–1.26		0.021
Yoshihara et al. [170]	PFS	High vs. low score	1.64	1.27–2.13		0.0001
Denkert et al. [37]	OS	Low vs. high score	1.7	1.1–2.6		0.021
Crijns et al. [32]	OS		1.94	1.19–3.16		0.008
Netinatsunthorn et al. [115]	RFS	Yes vs. no WT1 staining	3.36	1.60–7.03		0.0017
Spentzos et al. [148]	OS	Resistant vs. sensitive	3.9	1.3–11.4	41 months vs. not reached	$< 0.001^{\dagger}$
Raspollini et al. [128]	OS	No vs. yes COX-2 staining	0.23	0.06–0.77		0.017
Spentzos et al. [147]	OS	High vs. low score	4.6	2.0–10.7	30 months vs. not reached	0.0001

† Calculated value,
HR: Hazard Ratio,
OS: Overall Survival,
RFS: Relapse Free Survival,
PFS: Progression Free Survival,
CI: Confidence Interval

an independent data set was much less common. As may be seen from Appendix Table A.6, the success of many models was judged using the significance of covariates including the gene signature in the model. It is likely that this model was not applied to external data sets due to subtleties in what the model predicts when compared to methods such as linear regression. Whereas in linear regression the survival times are predicted directly, Cox proportional hazards regression predicts hazard ratios. Royston and Altman [133] developed techniques for the external validation of Cox proportional hazards models by application to an independent data set. These rely on having at least the weights of the variables included in the linear predictor, and ideally the baseline survival function. The first allows the assessment of the discriminatory power of a model, whereas the second is also required to allow the calibration of the model to be assessed. Royston and Altman [133] are of the opinion that the inclusion of a log-rank test p-value is not informative, due to the irrelevance of the null hypothesis being tested, and hence this should not be considered when judging model performance. An alternative to the log-rank test to compare survival between groups would be time-dependent ROC curves [64].

Failure to Predict

Of the studies identified by this review, some models failed to achieve significant predictive ability. These include Lisowska et al. [90], Vogt et al. [161] and Brun et al. [22]. Of these papers, Vogt et al. [161] and Brun et al. [22] both considered small numbers of genes when constructing their models. It is possible then that these models failed because no informative genes were considered. Conversely, Lisowska et al. [90] applied their modelling technique to over 47000 genes using 127 patients. It is therefore a possibility that genes were selected by their model purely by chance rather than due to true explanatory ability. This model was tested using an independent data set. When the model was applied to this data set it performed poorly, suggesting that the genes chosen did not generalise to the second cohort of patients. Neither Vogt et al. [161] nor Brun et al. [22] reported measuring the precision or accuracy of the gene expression measurements. Lisowska et al. [90] used RT-PCR to measure the expression of 18 genes from the microarray, but the RT-PCR measurements were carried out on a separate set of samples and hence are not useful when considering accuracy. It is therefore unknown whether the gene expression measurement techniques applied by these studies were sufficiently accurate.

2.4 Discussion

The papers identified as part of this review tackled the important issue of chemoresistance and survival prediction in ovarian cancer via gene or protein expression. The concept of identifying gene signatures is popular, and requires careful handling to extract the information required for this to be successful. It was observed that of the many different tissue preservation techniques applied, the most common were fresh-frozen and formalin fixed, paraffin embedded tissue. Due to the high quality expression measurements that may now be achieved with FFPE tissue, this appears to be the most appropriate choice for research intended to translate into a clinical setting.

It was found that the majority of the studies included in this review were heterogeneous with respect to the histological type of the patient cohort. This suggests that, due to the differing response of different types of ovarian cancer to chemotherapy, the gene signatures may be identifying different pathways and mechanisms. However, it should also be noted that although 27 of the 42 studies were heterogeneous, 12 of these consisted of greater than 80% serous samples. Therefore, for these studies the inclusion of multiple histological types is likely to have less effect on the gene signature and mechanisms highlighted could be expected to occur in serous ovarian cancer. It would be advisable for future studies to include histological type and grade as model features.

The majority of studies identified by this review attempt to classify patients into groups with different characteristics, for example 'poor prognosis' and 'good prognosis' or 'chemosensitive' and 'chemoresistant'. However, variables such as response to chemotherapy and prognosis are rarely so well separated into classes; they are by nature continuous variables. Altman and Royston [11] are clear that dichotomising continuous variables into categories (such as high-risk vs. low-risk) should be avoided, as it results in loss of information and may lead to underestimation of variation and the masking of non-linearity. Arbitrary choices of cutoff values may further obscure the situation, when the original continuous variable could serve the same purpose in many models. In terms of a clinical test it therefore may be more appropriate to apply alternative techniques, such as various types of regression, to obtain a real valued prediction of patient outcome.

It was noted that the metrics reported as measures of predictive ability vary between studies. These vary in the amount of information conveyed and hence care should be taken to use metrics that fully describe the model. Sensitivity and specificity are commonly reported for classification techniques and, together with the

numbers of patients in each class in the data set, allow the probabilities of a patient having the condition of interest given that they have tested positive or negative. It is the ultimate aim of most classification studies to obtain these probabilities, as it allows the predictive ability of the test to be assessed and the applicability of the test to be evaluated. Of the studies reporting sensitivity, specificity and related information, the best predictive ability was achieved by Hartmann et al. [63] and the worst by Helleman et al. [66]. It is important to note that from the sensitivity and specificity the model by Helleman et al. [66] does not appear to be any worse than some of the others, but these probabilities incorporate the prevalence of the condition of interest in the test population. It would therefore be highly informative to recalculate these probabilities using the prevalence of the condition in the population of ovarian cancer patients. Since some of the test populations were not representative of the overall population (having so called 'spectrum bias'), this would give a much more reliable indication of the predictive ability of the models in a clinical setting.

One of the main aims of the studies identified was to obtain a 'gene signature', the expression of which can explain and predict the response in the patient. To this end, the majority of the papers (32 of 42) provided full or partial list of the genes selected by the modelling process. An analysis of these gene signatures resulted in the conclusion that the signatures were very dissimilar, with the most commonly selected gene appearing in only four papers. 93.53% of genes were selected by only one paper. This seems to indicate that the gene signatures identified were not based on underlying cellular processes, or at least that the processes being highlighted were not the same across the papers. It should be noted that many of the studies used cohorts of patients who were heterogeneous in terms of chemotherapy treatment and, due to the development of resistance to chemotherapy via gene expression changes, this may affect the genes found to be explanatory. It may be that several gene signatures from sub-populations of patients treated with different drugs are combining and hence reducing the predictive ability of the models.

In order to assess the biological relevance of the genes selected for the gene signatures, gene set enrichment analysis was carried out. This technique is used to highlight processes and pathways that are over-represented in the gene signature compared to the set of all genes for the relevant organism, as defined by the database in use. For the purposes of this review, two groups of studies were considered: those where the patients were treated with platinum and taxane, and those where the patients were treated with other platinum based treatments. These groups were selected due to the low numbers of studies using a single treatment option. For

example, there were no studies considering platinum, taxane or cyclophosphamide as single agents. Following the analysis, 30 KEGG terms were returned for each group. Of these, each list comprised of approximately half cancer related terms. Of these the majority were processes often up- or down-regulated in cancer cells, such as proliferation, apoptosis, and motility and metastasis [62]. It is unclear whether the change in regulation of these processes is further altered in response to specific chemotherapy treatments. However, one process worthy of additional consideration is DNA repair. DNA repair is known to be an important mechanism in cancer, both though cancer development when down-regulated or mutated [119] and resistance to DNA damaging chemotherapy when up-regulated [96]. Therefore, the strong presence of DNA repair terms may suggest the presence of platinum resistance pathways in the gene signatures. Although the combined gene signatures appear not to include predictive chemotherapy-specific information, they may be capable of providing prognostic information. It is also thought that some studies, such as Glaysher et al. [54], may include genes relevant to additional chemotherapy-specific processes which are 'drowned out' when combined with other signatures.

2.5 Conclusions

It is clear that the prediction of response to chemotherapy in ovarian cancer is an ongoing research problem that has been attracting attention for many years. However, although many studies have been published, a clinical tool is still not available. Progress within the field suggests that the development of a predictive model is possible.

There is great variability between the approaches and success of existing studies in the literature, and there have been very high levels of variation in the genes identified as explanatory. If more care is taken when selecting the patients for inclusion to control for treatment history, these gene signatures may be simplified and models able to predict response to treatment may be developed.

Chapter 3

Gaussian processes for survival data with right-censoring: Theory

In this chapter three Gaussian process models are presented, capable of handling right-censored survival data. Here the right-censored time-to-event values, produced if a subject leaves a study, are modelled as partially missing outcomes, in terms of a set of known covariates, using Gaussian processes to place priors on the space of all functions relating the covariates and outcome. Due to their attributes as Gaussian processes, these models are flexible and probabilistic, making them ideal for handling noisy and complex data sets such as biomedical data.

3.1 Introduction

The analysis of survival data is an important aspect of medical research. Survival time is defined as the period from some starting point to the time at which a pre-defined event occurs. In a medical context, this event is often death, but could equally be disease relapse or symptom occurrence. The analysis of survival data often consists of identifying biomarkers capable of predicting survival time, either prognostically or predictively.

Survival data are often affected by censoring. Right-censoring occurs when a patient leaves a study, for a reason other than the pre-defined event, so the recorded time is a lower bound on the survival time rather than an actual survival time. This censoring reduces the amount of information available in the data set as the outcome of some patients will not be known. For survival data without censoring, simpler

statistical methods may be applied for analysis, e.g. regression, but in the presence of censoring, methods must be able to account for this missing information or the conclusions will be biased towards shorter survival times. Popular existing models for survival analysis include Cox proportional hazards regression [29], accelerated failure time models [163] and more general models modified for survival data, such as generalised boosted models [138] and Random Survival Forest models [73].

In cancer research, in addition to clinical patient data such as age, stage and grade, large amounts of molecular data are now often available. For example The Cancer Genome Atlas (TCGA) project generated a range molecular data types including gene expression measurements, somatic mutations and copy number variations [70]. The information available in these data is likely to be relevant to patient survival as it represents the processes underlying the action and progression of the disease. These data are usually high dimensional and relationships between measured variables are expected to be complex and non-linear. The models applied to them therefore need to be able to accurately capture and represent this complexity.

Currently in the literature, the majority of studies only apply basic analyses to these data, with by far the most common model used being Cox proportional hazards regression [29]. This model defines the hazard of a patient at time t to be $\lambda(t|\mathbf{x}) = \lambda_0(t)\exp(\mathbf{x}\boldsymbol{\beta})$, where $\mathbf{x}$ are the covariates for the patient, $\boldsymbol{\beta}$ are the model coefficients, and $\lambda_0(t)$ is the baseline hazard. This model does not require the baseline hazard to be known, and the model is simply fitted for $\boldsymbol{\beta}$. Due to the lack of knowledge concerning the baseline hazard, predictive ability is often reported as hazard ratios - the ratio of hazards for two groups - as the baseline hazard cancels.

This model makes assumptions about the data and relationships to the target value, one of which is the 'proportional hazards' assumption. As may be seen from the form of the hazard above, any time dependence is present via the baseline hazard. The proportional hazards assumption therefore requires that the effect of a feature on the hazard function is multiplicative and, unless specified otherwise, constant over time. This assumption is not necessarily true; features such as tumour grade and hormone receptor status have been found to have time-varying effects [16, 15]. In the case of molecular measurements, when analysing a lung cancer gene expression data set Dunkler et al. [39] identified genes with both converging or diverging hazards. These observations suggest that, although the Cox proportional hazards model is effective under the required assumptions, alternative models are required in the case of more complex data.

Here Gaussian processes have been chosen as the basis for a survival model due to their flexibility. These models place priors on the space of functions relating

covariates to outcome, and hence are flexible and probabilistic, allowing them to effectively deal with noisy and complex data sets. In this way, very few assumptions are made about the form of the relationship between data and target, allowing the method to be applicable in many contexts.

Censored survival times may be considered to be missing, but with information known about their minimum possible value. In order to apply Gaussian processes to survival data, three models have been developed to learn values for the censored survival times, incorporating information from the censored samples, resulting in a training set composed of both censored and uncensored samples. Hence these models allow many more samples to be retained, using the data set to its best advantage.

3.2 Gaussian process regression

Gaussian process regression is a well researched topic, providing a comprehensive toolbox of methods suitable for model fitting in many varied contexts. A Gaussian process is a collection of random variables for which the joint distribution between any finite subset is Gaussian. A Gaussian process is completely defined via its covariance and mean functions, $k(\mathbf{x}, \mathbf{x}')$ and $m(\mathbf{x})$, and a random function $f(\mathbf{x})$ may be written $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ [127].

It is assumed that for some unknown, underlying function f, $y = f(\mathbf{x}) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$.

It may then be observed that for any finite set of points X, the Gaussian process on the latent function values $\mathbf{f}$ defines the joint distribution

$$p(\mathbf{f}|X) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, K) \tag{3.1}$$

where $\boldsymbol{\mu} = \mathbf{m}(X)$ and $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$ [104].

Using a Gaussian likelihood and prior, the marginal posterior probability distribution of the noisy target values may be calculated as

$$p(\mathbf{y}|X) = \int p(\mathbf{y}|\mathbf{f}, X)p(\mathbf{f}|X)\,\mathrm{d}\mathbf{f} \tag{3.2}$$

$$= \int \mathcal{N}\left(\mathbf{y}|\mathbf{f}, \sigma_n^2 I\right)\mathcal{N}\left(\mathbf{f}|\boldsymbol{\mu}, K\right)\mathrm{d}\mathbf{f} \tag{3.3}$$

$$= (2\pi)^{-\frac{n}{2}}|K + \sigma_n^2 I|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^\top (K + \sigma_n^2 I)^{-1}(\mathbf{y} - \boldsymbol{\mu})\right) \tag{3.4}$$

.

Model training involves inferring the hyperparameters θ that maximise the

marginal posterior probability distribution.

The joint distribution of the noisy observed target values and the latent function values for the test data is [104]:

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{m}(X) \\ \mathbf{m}(X_*) \end{pmatrix}, \begin{pmatrix} K(X,X) + \sigma_n^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{pmatrix} \right) \tag{3.5}$$

Following training on training targets and data, $\mathbf{y}$ and X, predictions $\hat{\mathbf{y}}_*$ may therefore be made using unseen data X_* as such [127]:

$$\hat{\mathbf{y}}_* = \mathbf{m}(X_*) + K(X,X_*)^\top (K(X,X) + \sigma_n^2 I)^{-1}(\mathbf{y} - \mathbf{m}(X)) \tag{3.6}$$

$$\mathbb{V}[\mathbf{y}_*] = K(X_*,X_*) - K(X,X_*)^\top (K(X,X) + \sigma_n^2 I)^{-1} K(X,X_*) + \sigma^2 I \tag{3.7}$$

The choice of the mean and covariance functions may be guided by the structure of the data. For Gaussian processes applied here and in Chapter 4, unless stated otherwise the chosen forms are the zero mean function and the squared exponential covariance function:

$$m(\mathbf{x}_p) = 0 \tag{3.8}$$

$$k_y(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \exp\left(-\tfrac{1}{2l^2} |\mathbf{x}_p - \mathbf{x}_q|^2\right) + \sigma_n^2 \delta_{pq} \tag{3.9}$$

where σ_f^2, σ_n^2 and l are function variance, noise variance and length scale hyperparameters, and $\mathbf{x}_p$ and $\mathbf{x}_q$ are the data for the pth and qth samples respectively.

As seen above, for the squared exponential covariance function seen in Equation 3.9, the hyperparameters are the function variance, noise variance, and length scale. These hyperparameters may be considered to represent different aspects of the data. Firstly, the function variance hyperparameter determines how far the function may vary from the mean. Secondly, the noise variance hyperparameter determines the noise level in the model. Lastly, the length scale hyperparameter represents the characteristic length scale of the model, that is, the speed at which y may vary with $\mathbf{x}$. These hyperparameters therefore determine the characteristics of the functions described by the Gaussian process. Similarly, other covariance and mean functions have their own associated hyperparameters, but these are not considered here.

3.3 Gaussian processes for survival

Three Gaussian process models have been developed for right-censored survival times. Here censored survival times are considered to be missing values for which

lower bounds are known. Given inferred missing values, the task is then reduced to regression using a full data set with no missing survival times. This approach utilises the information contained in both the censored and uncensored samples to impute an underlying, non-censored data set.

GPS1 - In this model censored target values are assumed to be unknown parameters, which are inferred subject to a known minimum target value for each censored case.

GPS2 - This model involves the same censored target learning as GPS1, but censored target values are assumed to have higher noise variance than uncensored samples. For these samples an extra hyperparameter is incorporated to model this uncertainty as an additional source of Gaussian noise.

GPS3 - As GPS2, except the additional source of Gaussian noise is taken from the predictive distribution for each censored target value, rather than being learned as an additional hyperparameter. This results in different additional noise contributions for each of the censored samples.

3.3.1 GPS1

The training targets for censored samples are considered to be missing, with the censored value being the minimum possible value, and hence the aim of this model is to infer these missing values as parameters and learn updated values for the censored training targets using the whole training set. In this way the censored samples may still be included in the training set of the model, allowing prediction to incorporate information from all censored and uncensored samples.

The training data set S consists of, for each sample, the data $\mathbf{x}$, such as clinical or molecular measurements, the target y, the time to event such as survival time to death, and the event flag e, where 1 denotes event occurred and 0 that it did not. The training set of targets and data, S, is partitioned by two subsets consisting of censored and uncensored samples, $S_{e=0}$ and $S_{e=1}$. These two subsets are treated differently by the algorithm. The uncensored samples remain unchanged throughout the algorithm, as their final event times are known, and for the censored samples updated target values are learned.

A Gaussian process is trained using the training targets and data S and used to predict $\hat{\mathbf{y}}_{e=0}$, the new learned values corresponding to the censored training targets $\mathbf{y}_{e=0}$. In this way, the new target values for the training set incorporate information from all of S and hence the final model will contain more information than if the model was only trained on $S_{e=1}$.

After prediction, the new training target values are considered in the context

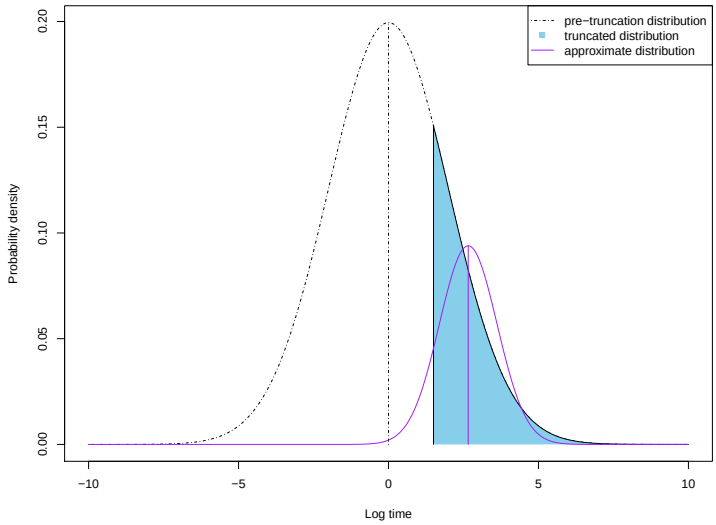


Figure 3.1: Without adjustment, imposing censoring on the predicted distribution results in a sharp cut-off, whereby all values below the censored value must be assigned the censored value. Instead, a truncated normal distribution is calculated, providing new mean a and variance values. The predicted distribution is then a normal distribution approximating this truncated distribution, by having the same mean and variance.

of the original censored target values. As GP predictions are distributions fully specified by the mean and the variance, it is useful to consider the censoring as truncation of this distribution. The expected value and variance of this normal distribution $A \sim N(\hat{\mu}, \hat{\sigma}^2)$ truncated at $A = c$, where c is the corresponding known censored target value $y_{e=0}$, are then calculated as

$$\mathbb{E}(A|c < A) = \hat{\mu} + \hat{\sigma}\lambda(\alpha) \tag{3.10}$$

$$\mathbb{V}(A|c < A) = \hat{\sigma}^2(1 + \delta(\alpha)) \tag{3.11}$$

where $\alpha = \frac{c-\hat{\mu}}{\hat{\sigma}}$, $\lambda(\alpha) = \frac{\phi(\alpha)}{1-\Phi(\alpha)}$ and $\delta(\alpha) = \lambda(\alpha)(\lambda(\alpha) - \alpha)$. $\phi(\alpha)$ are $\Phi(\alpha)$ are the probability density function and cumulative distribution function of the standard normal distribution respectively, both evaluated at α. The expected value of this censored distribution is therefore the predicted target value $\hat{y}_{e=0}$. Illustration may be seen in Figure 3.1.

This results in an updated target value for each censored training sample, creating an updated training set consisting of uncensored samples and previously censored samples with predicted non-censored target values.

3.3.2 GPS2 and GPS3

The first model, **GPS1**, simply uses the means of the predicted distributions as the replacement censored target value. However, this method leads to the underestimation of the noise variance hyperparameter, which may be seen in Appendix Figure B.1. In order to combat this two alternative approaches were developed, GPS2 and GPS3.

GPS2 involves including an extra hyperparameter to represent this additional variance for the censored samples, σ_c. This additional variance is included only when considering censored samples, as

$$k_y(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \exp\left(-\tfrac{1}{2l^2}\left|\mathbf{x}_p - \mathbf{x}_q\right|^2\right) + H_{pq}, \tag{3.12}$$

where H is the diagonal noise variance matrix, and

$$H_{pp} = \begin{cases} \sigma_n^2 + \sigma_c^2 & p \text{ where } e = 0 \\ \sigma_n^2 & p \text{ where } e = 1 \end{cases}.$$

This method learns a single hyperparameter, σ_c^2, used for all censored training samples.

GPS3 utilises the variance output by the model when the new training targets are predicted. This provides a variance per censored training sample, σ_c, which are included in the model as

$$k_y(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \exp\left(-\tfrac{1}{2l^2}\left|\mathbf{x}_p - \mathbf{x}_q\right|^2\right) + H_{pq}, \tag{3.13}$$

where H is diagonal, $H_{pp} = \sigma_n^2 + (\boldsymbol{\sigma}_\mathbf{c}^2)_p$,

$$\boldsymbol{\sigma}_\mathbf{c}^2 = \begin{cases} \mathbb{V}[A|A \sim N(\hat{\mu}, \hat{\sigma}^2), A > c],\ c \in \mathbf{y}_{e=0} & p \text{ where } e = 0 \\ 0 & p \text{ where } e = 1 \end{cases}$$

and $(\boldsymbol{\sigma}_\mathbf{c}^2)_p$ is the pth element of $\boldsymbol{\sigma}_\mathbf{c}^2$.

3.4 Initialisation

Also implemented are a hyperparameter pre-learning step and hyperpriors. The censored targets are simply initialised using a GP regression model trained on a proportion of the uncensored samples to provide initial guesses for $\hat{\mathbf{y}}_{e=0}$ and the hyperparameters. Identifying a sensible starting point for the hyperparameters may be used to reduce running times and may aid in the identification of the informative model regime, rather than very high or very low noise regimes.

Prior to model fitting, the target values are log transformed. This changes the

possible range of values from $(0, \infty)$ to $(-\infty, \infty)$, and will prevent negative survival times being considered.

Hyperpriors are implemented to provide information with regards to the hyperparameter values, using Gamma distributions on the log of the hyperparameters. After normalisation of each feature to $\mathcal{N}(0, 1)$ it is reasonable to expect σ_n^2 and σ_f^2 to be below 1, and for l to be smaller than the range of the data. Therefore, the hyperpriors were chosen to be $\sigma_n^2 \sim \mathcal{G}(k = 2, \theta = 1)$, $\sigma_f^2 \sim \mathcal{G}(k = 2, \theta = 1)$ and $l \sim \mathcal{G}(k = 2, \theta = d)$ where d is the 5–95th percentile range of the data X.

Plots of the hyperpriors may be found in Figure 3.2.

3.5 Implementation

All versions of the model have been written in R [126]. The functions and scripts may be found at `https://github.com/kllloyd/Thesis`.

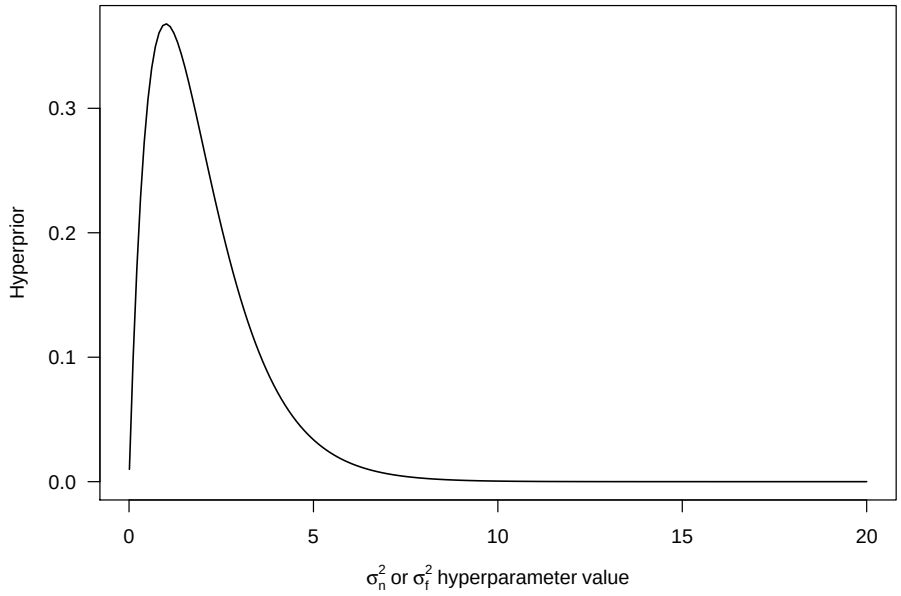
3.6 Inference

The algorithm applied for GPS1, GPS2 and GPS3 can be found in Algorithm 3.1.

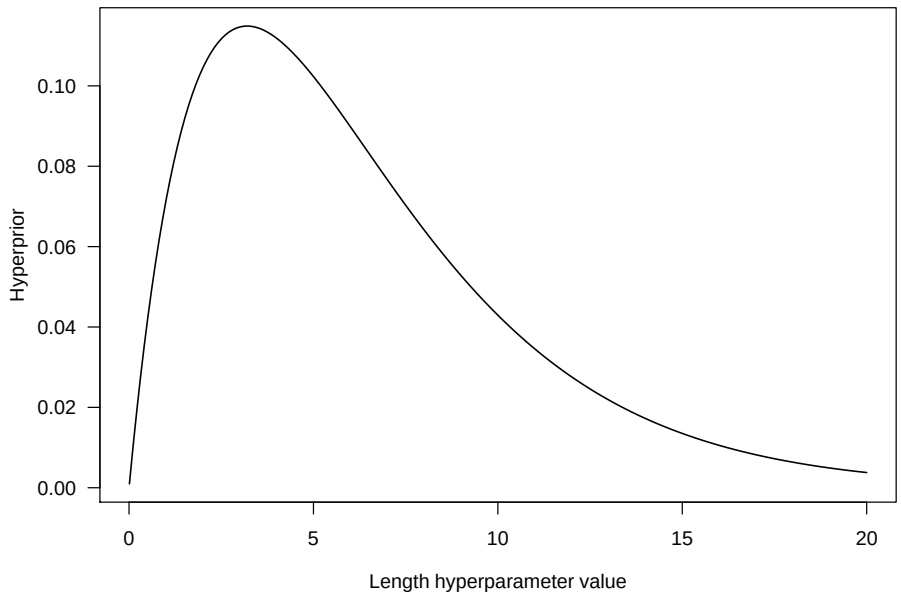
The original training set is the union of the censored and uncensored subsets, $S_{e=1} \cup S_{e=0}$, and consists of data X and uncensored and censored targets, $\mathbf{y}_{e=1}$ and $\mathbf{y}_{e=0}$. Using this training set, updated values for the censored targets are inferred, $\hat{\mathbf{y}}_{e=0}$. These replacement values are incorporated into the updated training set $S_{e=1} \cup \hat{S}_{e=0}$ and used to again predict new censored target values. This process is repeated until convergence of the log marginal posterior, L.

As with Gaussian process regression, training is carried out via optimisation of the log marginal posterior with respect to the hyperparameters σ_n^2, σ_f^2 and l. Here for all Gaussian process models optimisation is implemented using the R function optim, using the algorithm 'Nelder-Mead' [114].

For the Gaussian process for survival algorithms, hyperparameter and censored training target values are updated alternately, analogously to Expectation Maximisation. Following hyperparameter learning, updated values for the censored targets are inferred using the trained GP. Using the updated training set the hyperparameters are then learned again, followed by censored target value prediction. These steps are iterated until the convergence of the log marginal posterior.



(a) Example hyperprior for σ_n^2 and σ_f^2 hyperparameters.



(b) Example hyperprior for length hyperparameter. For this example, the 5–95th percentile range of the data set was 3.2.

Figure 3.2: Hyperpriors for use with GPS1, GPS2 and GPS3.

Algorithm 3.1: Gaussian process for survival

Data: $S_{e=1} \cup S_{e=0}$, S_*, tolerance, mean function m, covariance function k

Result: test targets $\mathbf{y}_*$

begin
- $\hat{S}_{e=0} \leftarrow S_{e=0}$
- $t \leftarrow 1$
- **while** targetValueChange > tolerance **do**
  - train $GP(m, k)$, training set = $S_{e=1} \cup \hat{S}_{e=0}$
  - predict new target values $\hat{\boldsymbol{\mu}}$ and variances $\hat{\boldsymbol{\sigma}}^2$, test set = $\hat{S}_{e=0}$
  - $\tilde{y} \leftarrow \mathbb{E}[A|A \sim N(\hat{\mu}, \hat{\sigma}^2), A > c]$, $c \in \mathbf{y}_{e=0}$
  - targetValueChange $\leftarrow |L_{t-1} - L_t|$
  - $\hat{\mathbf{y}}_{e=0} \leftarrow \tilde{\mathbf{y}}$
  - $t \leftarrow t + 1$
- predict values for test targets $\mathbf{y}_*$, test set = S_*

3.7 Illustration

The results of censored target learning using GPS1 may be found in Figure 3.3. For illustration purposes, a small synthetic data set was generated in 1 dimension (100 training samples, 50 test samples) and then censored non-informatively, with 40% of training samples being censored. Figure 3.3 shows training set target values pre- and post censoring, and post learning. These learned target values would later be used as the training set for test set prediction. It is clear that GPS1 learned the pre-censoring target values successfully, with good matching between pre-censoring and post-learning values. Of particular interest is the far right of the data range, where the Gaussian process is informed mostly by censored target values, the largest of which is providing information about the minimum value for the adjacent points.

The hyperparameters learned by this model may be seen in Figure 3.4. Although the fit may be seen to be good in Figure 3.3, the hyperparameters learned do not match the generating hyperparameters very well. This is likely due to the uncertainty introduced by the censoring; the noise introduced by the censoring may be interpreted as variation in the underlying hyperparameters, leading to variation in the possible values selected.

To illustrate the effect of repeated hyperparameter learning and censored training target prediction, Figure 3.5 shows a sequence of plots of intermediate states whilst training a GPS model, using a one-dimensional synthetic dataset. These plots were selected to be representative of early, middle and late stages of the process.

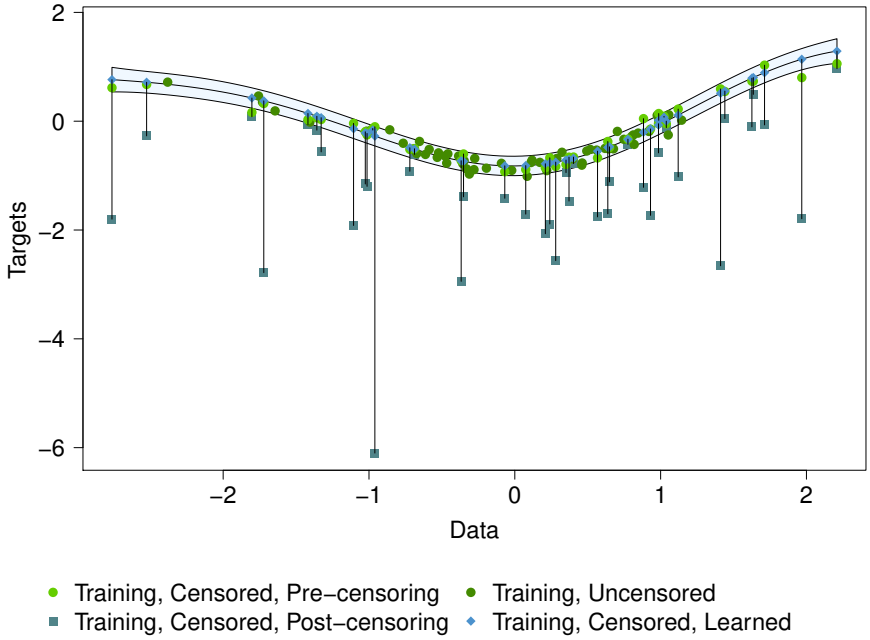


Figure 3.3: Plot of training set data against training set targets. Target values before and after censoring and after learning are shown. The mean $\pm 2\times$ standard deviation of predictions are also shown.

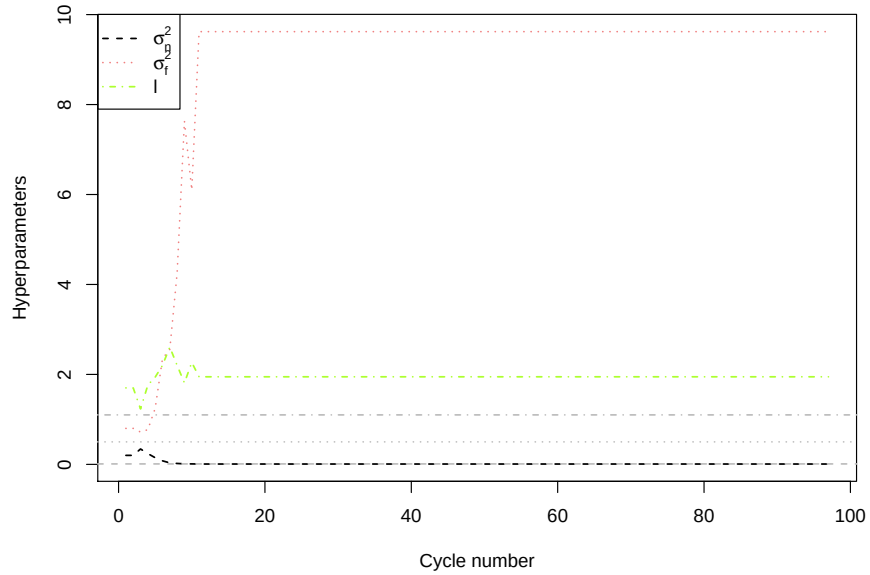
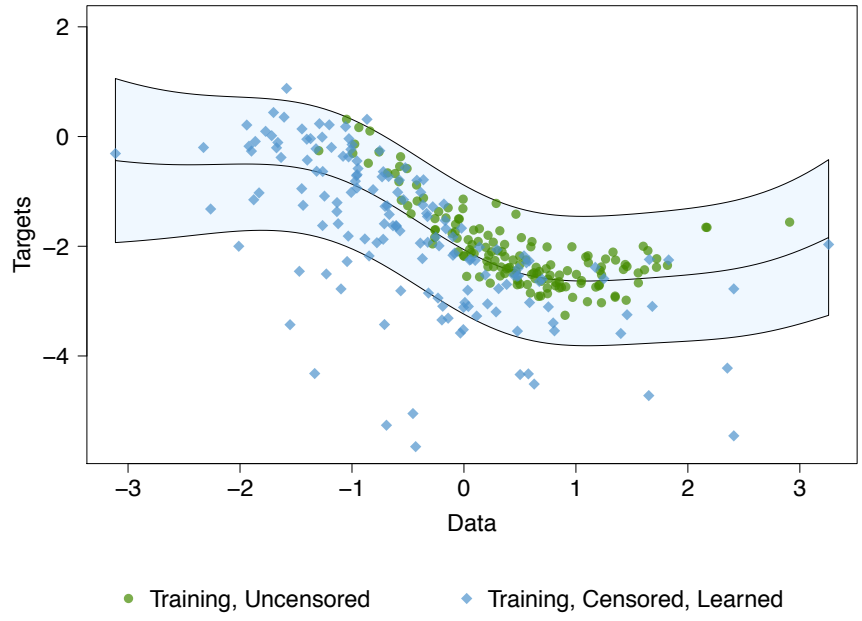
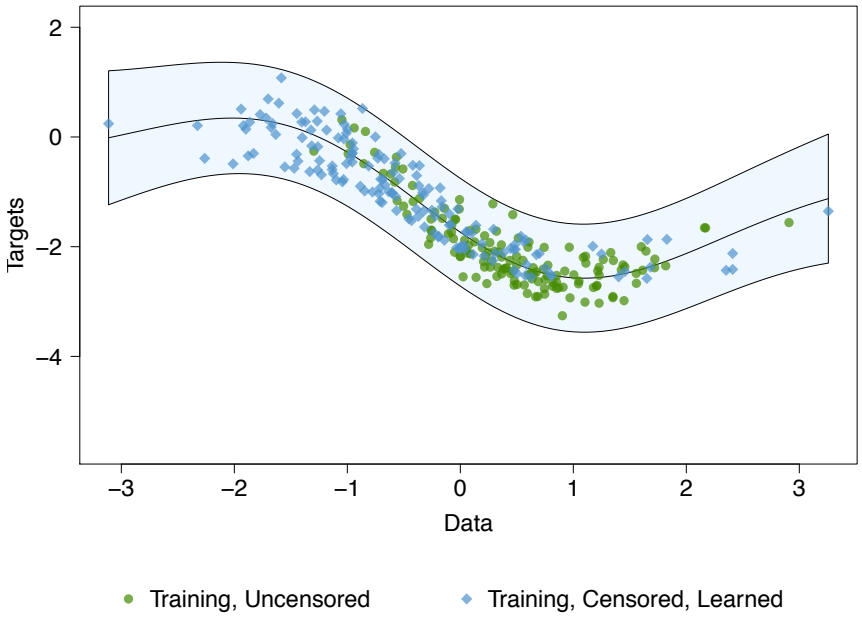


Figure 3.4: Plot of hyperparameters σ_n^2, σ_f^2 and l learned by GPS1. Generating hyperparameters are marked in grey using the corresponding line type.

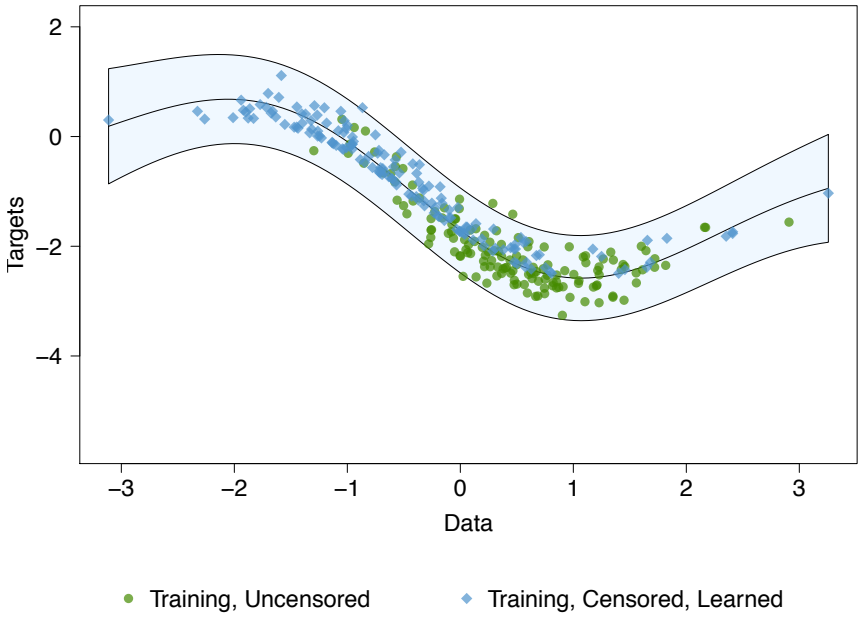


(a) Early

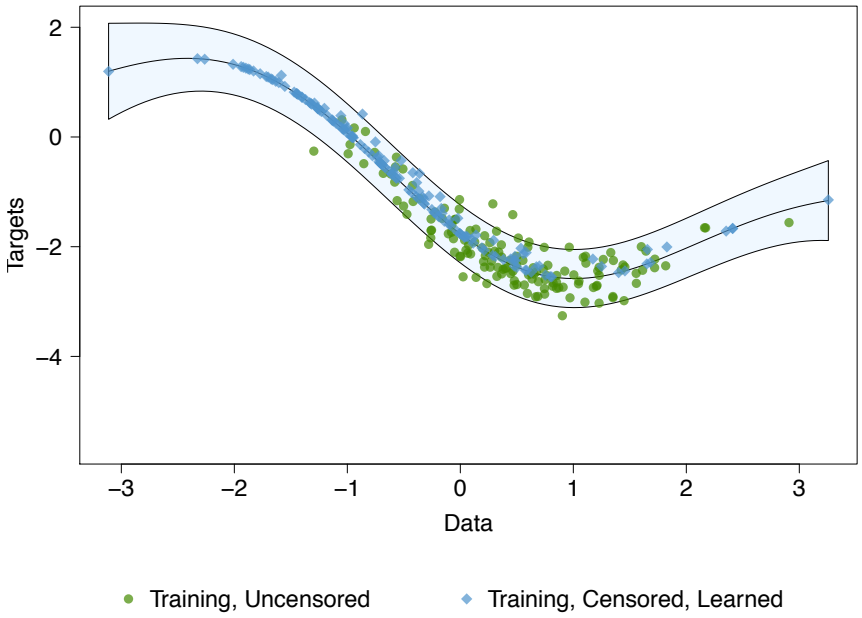


(b) Middle

Figure 3.5: A sequence of plots showing intermediate stages of the training set data and targets whilst training GPS models.



(c) Middle



(d) Late

Figure 3.5: A sequence of plots showing intermediate stages of the training set data and targets whilst training GPS models.

3.8 Conclusions

Three Gaussian process models have been developed for survival data, and infer survival times for censored training set samples. These models will allow the flexibility and non-linearity of Gaussian process models to be utilised when analysing complex medical data sets, such as gene expression and clinical covariates in cancer survival.

These models will be investigated in Chapter 4, using synthetic and real data sets, in parallel with other models for comparison.

Chapter 4

Gaussian processes for survival data with right-censoring: Numerical Experiments

Following the development of the Gaussian process models for survival data in Chapter 3, in this chapter these models have been applied to both synthetic and real data. Experiments using synthetic data allow model testing, and the application of the models to cancer molecular, clinical and survival data demonstrate their applicability to real, noisy, biomedical data. Other commonly used models were also applied to the same data for comparison.

4.1 Introduction

Chapter 3 detailed three Gaussian process models for right-censored survival data. These models were developed to extend the toolbox of Gaussian process methods applicable for regression to survival data, which would otherwise require censored samples to be removed from the training set.

As observed in Chapter 2, a relatively small range of models are commonly applied to survival data. In the context of comparison with the Gaussian process for survival models developed here, several models were selected, either due to popularity or suitability. These models are detailed in Section 4.2.6. One of the most popular is Cox proportional hazards regression [29]. This model makes the simplifying assumption that the hazard is proportional to the exponential of a linear combination of the model features [40]. The hazard for a subject at time t may be calculated as $\lambda(t|\mathbf{x}) = \lambda_0(t)\exp(\mathbf{x}\boldsymbol{\beta})$ where $\mathbf{x}$ are the covariates for the subject, λ_0

is the baseline hazard, and $\boldsymbol{\beta}$ are the fitted coefficients. When considering hazard ratios, a common measure of covariate predictive ability, this allows the baseline hazard function to be ignored as it cancels. However further analysis, such as using the model predictively, which requires the baseline hazard function to be known may be problematic, requiring additional assumptions to be made.

The Cox proportional hazards model may also be combined with elastic net regularisation [144]. For this method, conditions on the l_1 and l_2 norms are imposed on the model coefficients, $\boldsymbol{\beta}$, and these coefficients are then estimated as $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}(\|\mathbf{y} - X\boldsymbol{\beta}\|^2 + \lambda_2\|\boldsymbol{\beta}\|^2 + \lambda_1\|\boldsymbol{\beta}\|_1)$ where X is the covariate data and $\mathbf{y}$ are the target values. λ_1 and λ_2 are mixing weightings and are selected to prioritise either the ridge or the LASSO term. This technique allows the regression coefficients to be shrunk or set to zero, with the aim of minimising overfitting and identifying informative features when modelling high dimensional data. When applied in combination with a Cox proportional hazards model, this technique may be applied to survival data.

Similarly to the Cox proportional hazards model, the accelerated failure time (AFT) model relies on a baseline hazard, but covariates are also assumed to accelerate the time to failure [163] as the baseline hazard is also a function of the covariates: $\lambda(t|\mathbf{x}) = \lambda_0\left(\theta(\boldsymbol{\beta}, \mathbf{x})t\right)\theta(\boldsymbol{\beta}, \mathbf{x})$. $\theta(\boldsymbol{\beta}, \mathbf{x})$ is commonly chosen to have the form $\theta(\boldsymbol{\beta}, \mathbf{x}) = \exp(\mathbf{x}\boldsymbol{\beta})$

Gradient boosting machines (GBM) may also be applied to censored data. These methods utilise a series of additions to a weak model, chosen to minimise a loss function equivalently to gradient descent [46]. This forms an ensemble of weak models to create an overall strong learner. When Cox proportional hazards is chosen to be the base model, the GBM allows a more complex model to be built around this basic form.

Random Forest models may also be used to model survival data. These models create ensemble predictions using a large number of decision trees learned on random subsets of the available features [20]. Random forests for regression can only be applied to uncensored samples; however, a version for survival data has also been developed [73]. Due to its high reliability and generalisability, Random Forest is utilised for many different contexts, and hence is very important for comparison here. Similarly the Cox proportional hazards model, due to its popularity, is another important comparison method.

When developing new models, it is important to consider how they perform on a variety of data. When investigating the predictive ability of the model, data in regimes corresponding to real life data must be investigated. In this chapter,

synthetic data will be used to investigate the success of the Gaussian process for survival models in various regimes. Synthetic data is especially useful, due to known 'ground truth': the true values of the survival of the test samples, the pre-censoring survival times of the censored training samples, and the underlying hyperparameters used for data generation. These known values may then be compared to the learned and predicted values output by the model, allowing the fit to be assessed.

As these models were developed to deal with survival data, it is expected that a proportion of training samples will be censored. Due to underlying assumptions, for some models it is often assumed that censoring is non-informative, that is, there is no relationship between survival and probability of censoring. To implement non-informative censoring on synthetic data, this requires that samples are selected randomly for censoring, and that a randomly generated value is used to replace the survival time for that sample.

Following model testing on synthetic data, models should always then be applied to real life data. Due to the cleanness and simplicity of synthetic data, models may achieve higher predictive ability than would be seen in real data. It is important to note that model predictive ability may also be inflated by unnoticed bias in choices for the generative model, such as parameter and function selection. In this chapter, the models are therefore also applied to a selection of real data sets, consisting of survival, clinical and molecular measurements from cancer patients.

Code and functions to run the experiments detailed in this chapter are available at `https://github.com/kllloyd/Thesis`.

4.2 Methods

4.2.1 Synthetic data

Synthetic survival data were drawn from a Gaussian process with predefined hyperparameter values. A squared exponential covariance function and a zero mean function were used with hyperparameters set as in Table 4.1. These hyperparameters were selected to provide nonlinear trends, changing on a reasonable length scale, when considered per feature. Hyperparameters denoted with a * in Table 4.1 were varied during experiments.

Data, X, were generated uniformly over [0,8] then passed to the Gaussian process to produce target values, $\mathbf{y}$. Samples for censoring were selected randomly and the target, y, was set to a censored target value chosen from a truncated normal distribution, $y_c \sim \mathcal{TN}(\mu = 0, \sigma = 50, 0 < y_c < y)$, ensuring non-informative censoring. Pre-censoring target values were stored to allow comparison with predictions.

Table 4.1: Table of parameters for synthetic data

Experiment	Dim	# Training Samples	# Test Samples	% Censoring	# Repeats	Hyperparameters σ_n^2	σ_f^2	l	**m**
1	6	400	50	75	30	0.10	0.7	1.1	**0**
2	6	500	100	*	30	0.10	0.5	1.1	**0**
3	6	*	100	75	30	*	0.9	1.1	**0**

* Run for a selection of values, see Section 4.3.1

Samples were then divided between training and test sets randomly. Each feature was standardised to have mean 0, standard deviation 1. For experiments involving repeats the specified number of data sets were generated separately, equivalently to biological replicates. The number of samples, dimensionality, level of censoring and number of repeats for each synthetic experiment may be found in Table 4.1.

4.2.2 Yuan et al. [171] cancer data

The GPS models and comparison models were applied to real data, recreating the analysis applied by Yuan et al. [171] to publicly available molecular and clinical data from cancer patients. Yuan et al. [171] applied Cox proportional hazards and Random Survival Forest models, with the feature selection methods outlined below.

Core clinical and genomic/proteomic data, training and test data set splits and example code were obtained from the Synapse homepage of the project (accession number syn1710282, doi:10.7303/syn1710282). The example code provided by Yuan et al. [171] was used to recreate the methods used, with the code being used as provided when possible, and modifications minimised. The data sets consist of samples as seen in Table 4.2. This data is a curated version of that generated by The Cancer Genome Atlas (TCGA) [118, 116, 101, 117]. The six data types are:

- clinical
- somatic copy-number alteration (SCNA)
- DNA methylation (methyl)
- mRNA expression (mRNA)
- microRNA expression (miRNA)
- protein expression (protein)

The cancer types considered are:

- kidney renal clear cell carcinoma (KIRC)
- glioblastoma multiforme (GBMF)
- ovarian serous cystadenocarcinoma (OV)
- lung squamous cell carcinoma (LUSC)

Table 4.2: Table of sample numbers and data availability for each cancer type, data from Yuan et al. [171]

Cancer	# Samples	# Variables					
		Clinical	SCNA	methyl	mRNA	miRNA	protein
KIRC	243	4	69	16 484	20 203	795	166
OV	379	3	109	24 980	17 813	798	165
GBMF	210	3	106	24 980	17 813	533	-
LUSC	121	3	114	-	20 194	829	174

Cox proportional hazards and Random Survival Forest methods were applied as by Yuan et al. [171], using the R packages *survival* [155] and *randomForestSRC* [73] were used respectively. In addition, feature selection was applied prior to model training. Univariate Cox Proportional Hazards models were used to identify significant features and those with a p-value < 0.05 were retained. For the Cox Proportional Hazards model, a lasso penalty was also applied if more than five features were remaining. This feature selection was applied as recreated from Yuan et al. [171], and functions used may be found in the thesis code repository (see chapter Introduction, Section 4.1).

All models were run on the same 100 cross-validation training and test sets, as specified by Yuan et al. [171]. For the GP models, each molecular feature was standardised to have mean 0, standard deviation 1.

4.2.3 Tothill et al. [156] cancer data

A second real data set was obtained using the the R package *curatedOvarianData* [48]. This package consists of data from studies assessing aspects of ovarian cancer using gene expression and data has been prepared to provide a consistent method of accessing features across studies. Other clinical features are also available in many data sets. The data set with the largest number of samples, other than TCGA, was utilised.

This data set was GSE9891, generated by Tothill et al. [156]. Clinical features were selected to be grade, stage and age at diagnosis, as these features have an

Table 4.3: Table of sample numbers and data availability, data from Tothill et al. [156]

Data Type	# Samples	# Variables
Clinical	274	3
mRNA	285	19 816
Clinical + mRNA	274	19 819

acceptably low proportion of missing data. The available numbers of samples and dimensionality of each data type can be found in Table 4.3.

For this gene expression data set, subsets of genes were identified for use in the models, as the full data set was prohibitively large. Information on these two gene sets, Ovarian Cancer Gene Set (OCGS), and Systematic Review Gene Set (SRGS), may be found in Section 4.2.4 below. These gene sets are applied as a filtering step prior to model fitting for all models. Following gene name standardisation to the HUGO system [57], only features with names corresponding to those in the gene set were retained.

4.2.4 Gene sets

In the context of cancer data, studies and measurement techniques are becoming increasingly high dimensional. Whilst this provides more information per patient, data sets such as gene expression microarrays can be prohibitively large (~ 20000 features) and hence some feature filtering is useful prior to model fitting. Here two separate gene sets are considered, the Ovarian Cancer Gene Set (OCGS) and the Systematic Review Gene Set (SRGS). Both gene sets were developed using prior knowledge from published studies and are hoped to provide genes relevant to cancer survival.

The first gene set, OCGS, was derived from a study by Glaysher et al. [54] for which a custom TaqMan gene expression microfluidic array was developed to explore resistance to chemotherapy in ovarian cancer. The genes selected were biologically motivated and were selected to be associated with various functions relevant in cancer: apoptosis, proliferation, pumps and detox, and DNA repair. A small number of housekeeping genes were also included for comparison. The gene set OCGS comprises of 97 genes included in the TaqMan array.

The second gene set, SRGS, was developed using the results of the systematic review by Lloyd et al. [92], also detailed in Chapter 2. The systematic review investigated the literature concerning the prediction of patient response to chemotherapy in ovarian cancer via statistical methods. The study resulted in a list of genes found

to be predictive by various published studies. The genes in SRGS were chosen to be those selected as predictive by two or more papers. This gene set consists of 84 genes.

There is a small overlap of eight genes between the two gene sets: APAF1, BAD, FASLG, HSPD1, SLC29A1, ABCC1, ABCC2, ABCC6, ERBB2, ERBB3, TP53 and VEGFA.

The genes comprising OCGS and SRGS may be found in Tables 4.4 and 4.5. Both of these gene sets will be used for feature filtering with the Tothill et al. [156] cancer data set.

4.2.5 Gaussian process for survival models

The experiments here investigate the Gaussian process for survival models, as detailed in Chapter 3: GPS1, GPS2, and GPS3. These models are extensions of Gaussian process regression to apply to right-censored data and here will be applied to the synthetic and real data sets along with other, standard models for comparison. Unless stated otherwise, these models are applied with squared exponential covariance function and zero mean function. Starting hyperparameter values were either random or not equal to the generating hyperparameters.

As outlined in Chapter 3:

- **GPS1:** Gaussian process for survival data with no noise hyperparameter correction
- **GPS2:** Gaussian process for survival data with noise hyperparameter correction, implemented as an additional learned hyperparameter
- **GPS3:** Gaussian process for survival data with noise hyperparameter correction, implemented using the predicted variance of each training target prediction

These models will be subsequently referred to as GPS1, GPS2, and GPS3 respectively.

4.2.6 Comparison methods

Other methods were also applied to the same data for comparison. All models were applied in R. The models used, with R functions and corresponding packages, were:

- **AFT:** Accelerated failure time, survreg (*survival*, Therneau and Grambsch [155])

Table 4.4: Genes contained in OCGS

Apoptosis	Proliferation	Pumps and Detox	DNA Repair	Housekeeping
AKT1	APC	ATP7B	ATM	TYMS
AKT2	TUBB3	ABCG2	BRCA1	HPRT1
AKT3	PTGS2	CES1	ERCC1	HMBS
APAF1	EGFR	CES2	ERCC2	SDHA
BAD	ERBB2	NT5C2	MGMT	TBP
BAX	ERBB3	DPYD	MLH1	
BCL2	ERBB4	FPGS	MSH2	
BCL2L1	HIF1A	H2AFX	MSH6	
BID	MKI67	GCLC	RAD51	
CFLAR	CDKN2A	GCLM	TOP1	
FAS	CDKN1A	GSTP1	TOP2A	
FASLG	CDKN1B	SLC29A1	TOP2B	
HSPD1	TP53	SLC29A2	XPA	
HSPA1A	VEGF	ABCB1	XRCC1	
HSPA1L		ABCC1	XRCC5	
HSP90AA1		ABCC2	XRCC6	
HSP90AB1		ABCC3		
HSP90B1		ABCC4		
BIRC2		ABCC5		
IGF1		ABCC6		
IGF1R		ABCC8		
IGF2		MVP		
IGF2R		UMPS		
IGFBP1		RRM1		
IGFBP2		SOD1		
DNAJC15		TAP1		
MCL1		TAP2		
FRAP1		ABCB4		
NFKB1				
PIK3CA				
PTEN				
STAT3				
BIRC5				
BIRC4				

Table 4.5: Genes contained in SRGS

AADAC	CPE	FN1	MMP1	SNX7
ABCB1	CXCR4	FOXA2	MUTYH	SRC
ABCB10	CXCR7	GNPDA1	MYCBP	TCF15
ACTR3B	CYP51A1	GUCY1B3	NBN	TGFB1
AGR2	DAP	HDAC1	NCOA1	TIAM1
AKAP12	DFNB31	HDAC2	NDST1	TIMP1
ALDH9A1	DGKZ	HMGCS1	NFIB	TOP2A
ANXA3	EFNB2	HSPB7	PCF11	TP53
AOC1	EHF	IGFBP5	POLH	TRIM27
ARHGDIA	EPHB2	IL6	PSAT1	TUBB4A
B4GALT5	EPHB3	ITGAE	RBM39	VEGFA
BAX	ERCC8	LBR	RFC3	XPA
BRCA2	ETS1	LGR5	RPL36	YWHAE
CD38	FADS2	LRIG1	S100A10	ZMYND11
CES2	FASLG	LSAMP	SDF2L1	ZNF12
CHIT1	FGFBP1	MARK4	SIVA1	ZNF200
COL3A1	FILIP1L	MECOM	SLC1A3	

- **Cox PH:** Cox proportional hazards regression, coxph (*survival*, Therneau and Grambsch [155])
- **GBM:** Gradient boosting machine, gbm (*gbm*, Ridgeway [130])
- **Coxnet:** Cox proportional hazards with elastic-net penalisation, coxph (*survival*, Therneau and Grambsch [155]), glmnet (*glmnet*, Friedman et al. [45], Simon et al. [144])
- **RF:** Random Forest for regression with censored samples removed, rfsrc (*randomForestSRC*, Ishwaran et al. [73])
- **RSF:** Random Forest for survival, rfsrc (*randomForestSRC*, Ishwaran et al. [73])
- **GP:** Gaussian process regression, applied only to uncensored samples (i.e. the censored samples were discarded)
- **GPR1:** Gaussian process regression, applied to all samples, with censored training times replaced by the median uncensored value
- **GPR2:** Gaussian process regression, applied to all samples, with censored training times replaced by values drawn from $U(c, \max(\mathbf{y}))$, where $c \in \mathbf{y}_{e=0}$ is the corresponding censored target time and $\mathbf{y}$ are the training set survival times.

These models will be subsequently referred to as AFT, Cox PH, GBM, Coxnet, RF, RSF, GP, GPR1 and GPR2 respectively.

For each repeat or bootstrap the same data or subset of the data were used for all models.

4.2.7 Assessing predictive ability: concordance index

Model predictive ability was assessed via the concordance index, which compares the rank order of predictions and the rank order of the measured values and incorporates censoring [149]. Two survival times may be ordered if either of two cases are true: both times are uncensored, or the uncensored survival time is smaller than the censored survival time. For each pair of orderable times, the ordering of the predicted times is compared to the ordering of the known times and a contribution to the sum is made if the orders match. The total is normalised by the number of possible comparisons between pairs, $|\mathcal{E}|$.

$$\text{c index} = \frac{1}{|\mathcal{E}|} \sum_{t_i \text{uncensored}} \sum_{t_j > t_i} \mathbf{1}_{\hat{t}_i < \hat{t}_j} \tag{4.1}$$

where t_i and t_j are known survival times corresponding to samples i and j, $\hat{t}_i$ and $\hat{t}_j$ are predicted survival times similarly, $\mathbf{1}_{a<b}$ is the indicator function $\mathbf{1}_{a<b} = \begin{cases} 1 & a < b \\ 0 & \text{otherwise} \end{cases}$ and $|\mathcal{E}|$ is the number of pairs of times which may be ordered.

Here, for synthetic data, comparisons are carried out between the predicted target values and the uncensored targets stored before censoring was applied. When considering real data, the uncensored values are unknown and so the predicted target values are compared to the measured data, including the censoring.

The concordance index was calculated in R using the function concordance.index from the package *survcomp* [140].

Statistical tests were carried out to test for differences in model predictive ability. To accompany these results, Cohen's d effect size was also computed. This is a measure of the difference in means between two sets of data and is normalised by the appropriate measure of variance. For paired samples this takes the form:

$$d = \frac{\text{mean}(\mathbf{z}_1 - \mathbf{z}_2)}{\text{sd}(\mathbf{z}_1 - \mathbf{z}_2)}. \tag{4.2}$$

where $\mathbf{z}_1$ and $\mathbf{z}_2$ are paired predictions by two models and the functions calculating the mean and standard deviation are applied to the difference between the two

vectors.

Statistical tests were carried out in R using the function t.test, which is a base function. Cohen's d effect size was calculated in R using the function cohensD from the package *lsr* [113].

4.3 Results and Discussion

4.3.1 Synthetic data

Synthetic data were generated as described in the Methods section of this chapter. GPS1, GPS2, GPS3, GP regression applied only to uncensored samples, and the selection of six comparison models were applied to the data. All models were applied to identical data sets, with the same training-test split of samples. For each experiment the data parameters were as found in Methods Table 4.1.

Experiment 1

For comparison between models, Figure 4.1 shows the concordance index of predictions made by the various models when applied to the same 30 six-dimensional synthetic data sets consisting of 400 training samples and 50 test samples, with 75% censoring (see Methods Table 4.1 for details).

It is clear that AFT, Cox PH and GBM perform poorly, with GBM having very low predictive ability. This is likely to be due to the presence of complex non-linear relationships in the data, which, with the exception of GMB, these models cannot account for. Random Survival Forest does slightly better, but is still not performing well. Importantly, the Gaussian process for survival models GPS1, GPS2 and GPS3 have noticeably higher mean concordance index than the less flexible models. The spread in values is also marginally smaller, suggesting that these models train and predict more reliably. The Gaussian process model trained on only the uncensored samples (GP) has a lower mean concordance index, to be expected as the high level of censoring results in a small training set. For comparison, Gaussian process regression was also carried out according to GPR1 and GPR2, with censored training targets assigned new values before training. These methods can be seen to be better than non-informative, but the resulting bias in the training set causes the predictive ability of these methods to be inferior to removal of the censored samples in this context, outweighing the benefit of the inclusion of a greater number of samples. This suggests that the GPS models perform better than simple censored

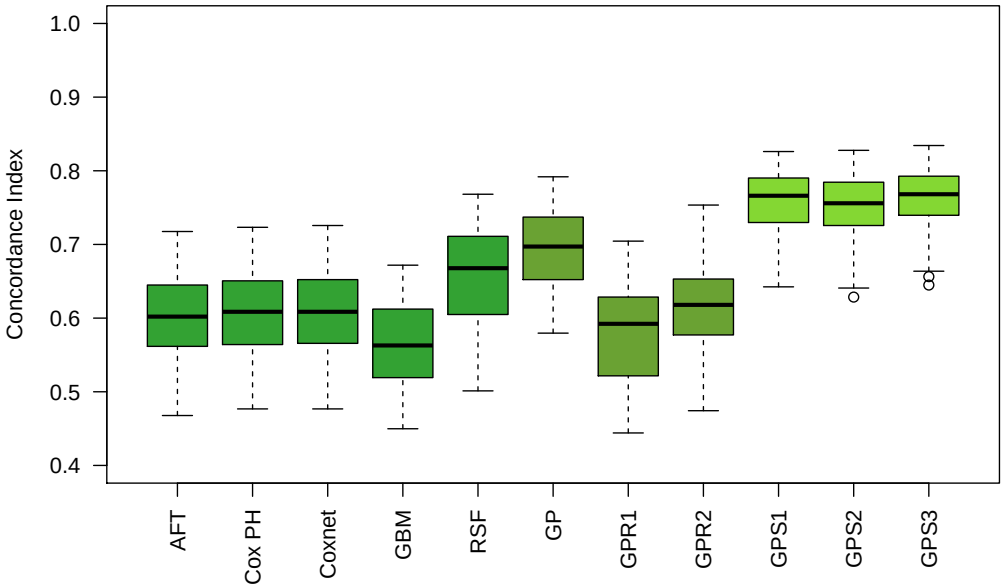


Figure 4.1: Experiment 1. Concordance index of test set predictions. Accelerated failure time, Cox proportional hazards, Cox proportional hazards with elastic-net penalisation, gradient boosting machine, Random Survival Forest, Gaussian process trained on only the uncensored samples, GPR1, GPR2, GPS1, GPS2 and GPS3 were applied to the same 30 synthetic data sets, generated using the same hyperparameter values. Boxplots show the median and first and third quartiles, with the whiskers marking 1.5 times the interquartile range from the box.

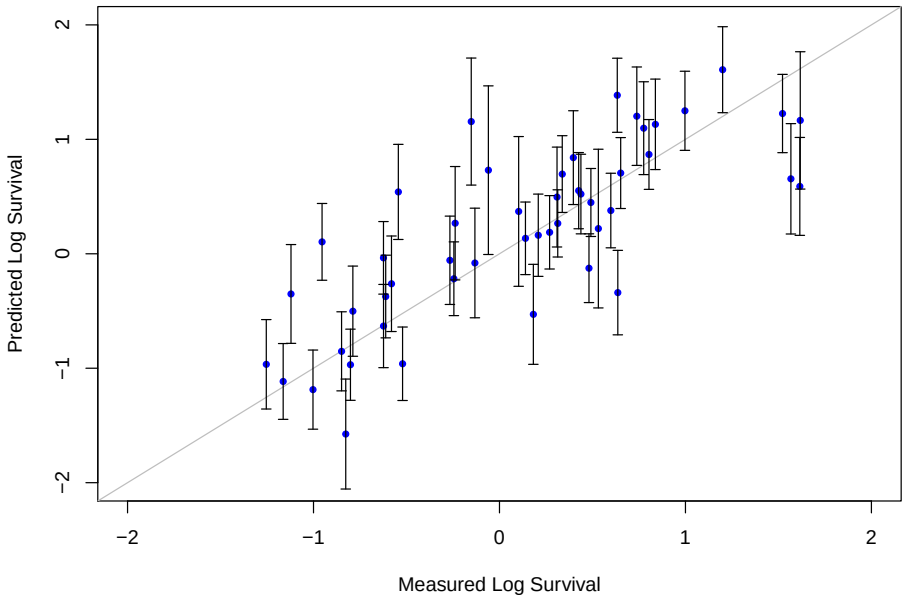


Figure 4.2: Experiment 1. Plot of pre-censoring versus predicted test target values for GPS3. Error bars show one standard deviation as calculated using the variance reported for each test sample by the model. $y = x$ line is marked in grey for reference. This replicate had a concordance index value of 0.8.

training target imputation.

As this experiment involved running the models on many replicates of the data, it may be informative to consider the predictions produced by one model. Figure 4.2 shows the log survival times predicted by GPS3 for a single replicate from Figure 4.1 against the pre-censoring test set target values. The error bars depict one standard deviation, as calculated using the variance reported for each prediction by the model. The $y = x$ trend may be seen clearly, though the error bars suggest that there is substantial uncertainty in these values on the part of the model.

Experiment 2

Figure 4.3 shows how predictive ability as measured by concordance index changes as the proportion of the training set that is censored is varied. As detailed in Methods Table 4.1, six dimensional data were generated with 500 training samples and 100 test samples. For each level of censoring, 30 data sets were generated to which all five models were applied.

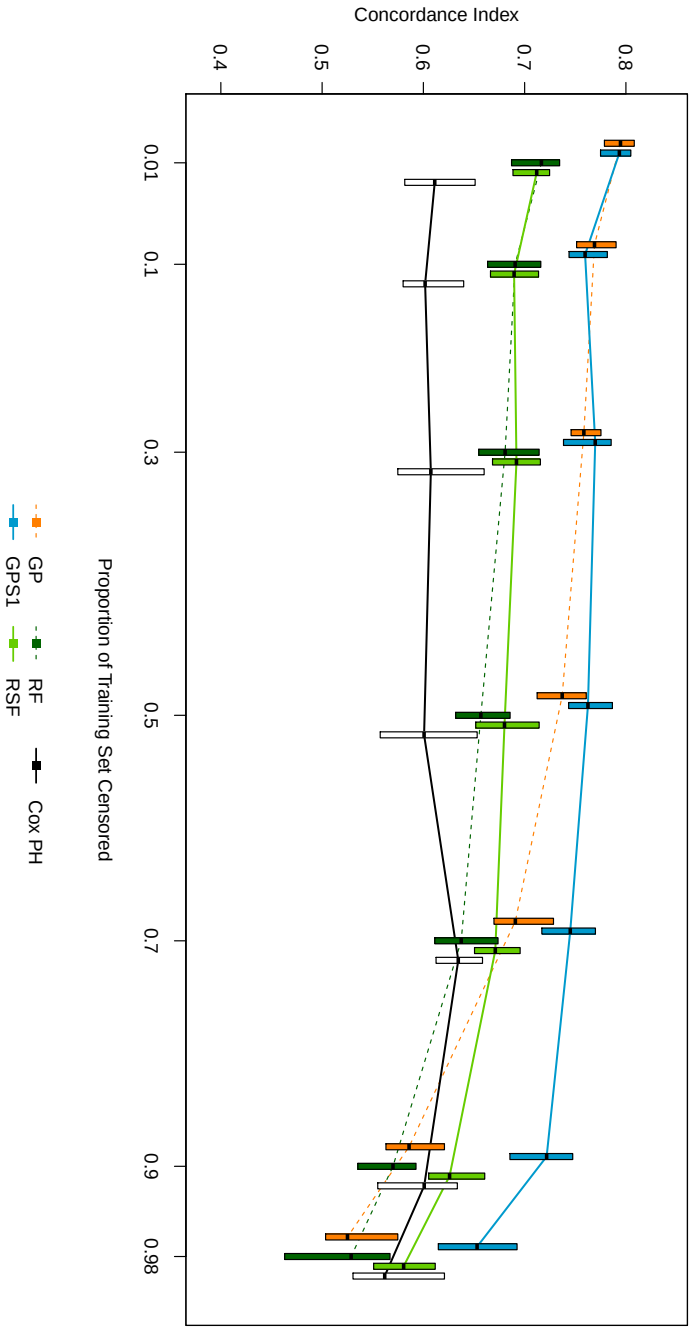


Figure 4.3: Experiment 2. Concordance index of test set predictions. GPS1, Gaussian process trained on only the uncensored samples, Cox proportional hazards, Random Forest trained on only the uncensored samples and Random Survival Forest were applied to the same 30 synthetic data sets, generated using the same parameter values, as the proportion of samples censored was changed. Boxplots show the median and first and third quartiles. For clarity, whiskers were not plotted.

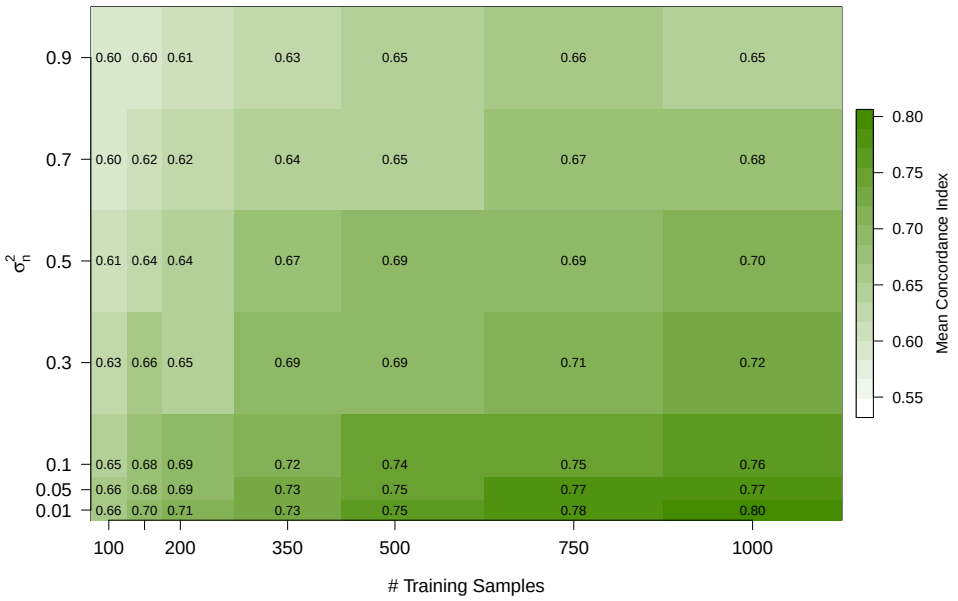
It may be seen that, although the Gaussian process model trained on only the uncensored samples is not able to predict accurately at high levels of training set censoring, GPS1 has a much greater ability to extract useful information from the censored samples in this situation. Both GP and GPS1 fare better than the Random Forest models at low censoring levels, though the survival models retain better predictive ability as the censoring level rises. Of the survival models in this experiment, Cox proportional hazards is the least successful.

Experiment 3

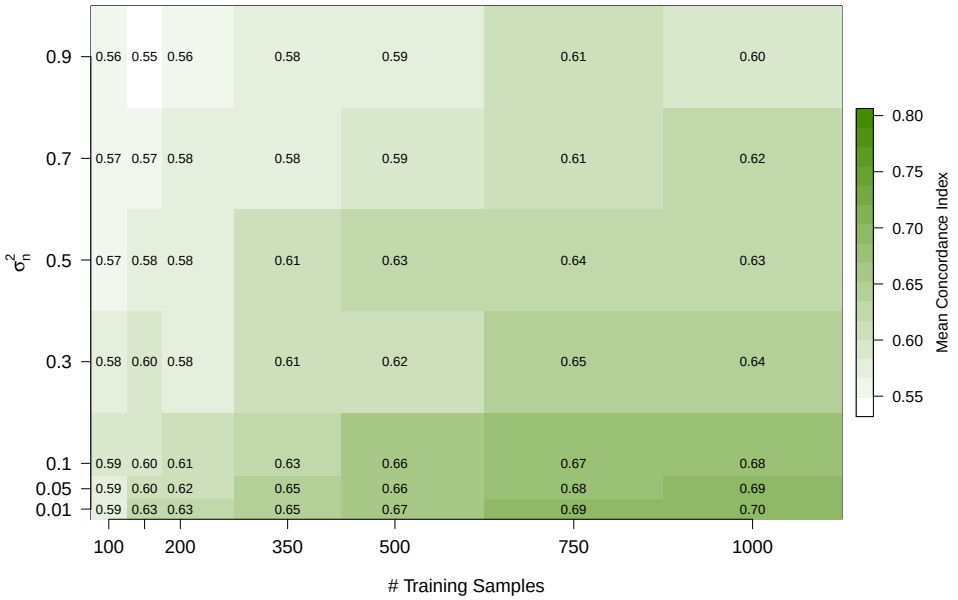
Medical data often contain high noise levels, and data sets are often limited in size, both due to sample availability and running costs. It is therefore of importance that the GPS models retain effectiveness in these circumstances. It was expected that the GPS models are likely to have their training and predictive ability affected by the data noise levels and the size of the training set. As these models work with censored data, the number of training samples is expected to be additionally important, as the number of uncensored samples will be smaller than the total number.

A series of data sets were therefore generated with varying values of the noise hyperparameter, σ_n^2, and the number of training samples, as detailed in Methods Table 4.1. 30 repeats were carried out for each cell. GPS3 was selected to be representative of the GPS models, as they were observed to provide similar results up to this point. GPS3 was chosen due to the inclusion of the additional noise term, unlike GPS1, and the observed lower computational time than GPS2. Random Survival Forest and Cox proportional hazards were selected as comparison models due to their frequent implementation and generally good predictive ability.

Figure 4.4a shows the mean concordance index for test targets from each generated synthetic data set, following prediction using the GPS3 model. It is clear that for low numbers of training samples and high noise levels the model does not perform well, but the concordance index values rise as these change. For low noise and larger numbers of samples the concordance index values achieved are respectable. For comparison, Figure 4.4b shows the same data fitted using the Random Survival Forest model. It is clear that this model was much less successful, with all cells showing lower mean concordance index than for the GPS3 model. When fitted using the Cox proportional hazards model, these values were lower still, with no cells having mean value above 0.65 (see Appendix Figure C.1).



(a) GPS3



(b) RSF

Figure 4.4: Experiment 3. Mean concordance index values as generating noise variance hyperparameter and number of training samples are varied. a) Model fitted was GPS3. b) Model fitted was Random Survival Forest.

4.3.2 Yuan et al. [171] cancer data

The Gaussian process models were also applied to data obtained from Yuan et al. [171], as outlined in Methods 4.2.2. These data are a curated version of the TCGA data for four cancer types: kidney renal clear cell carcinoma (KIRC), glioblastoma multiforme (GBMF), ovarian serous cystadenocarcinoma (OV) and lung squamous cell carcinoma (LUSC). Here, the clinical data alone for all four cancers and the molecular data with and without clinical data for KIRC were chosen to be run, due to Yuan et al. [171] achieving good results on these data sets. Cox proportional hazards and Random Survival Forest models were also run, as in Yuan et al. [171].

Using the training-test sample splits as defined by Yuan et al. [171], all models were applied to the data via 100-fold cross validation. This results in a set of test set predictions per model, per fold, and hence 100 concordance index values per model.

The resulting concordance index values for the models applied to the various cancers and platforms may be seen in Figures 4.5 and 4.6. The boxplots show the median and first and third quartiles, with the whiskers marking 1.5 times the interquartile range from the box.

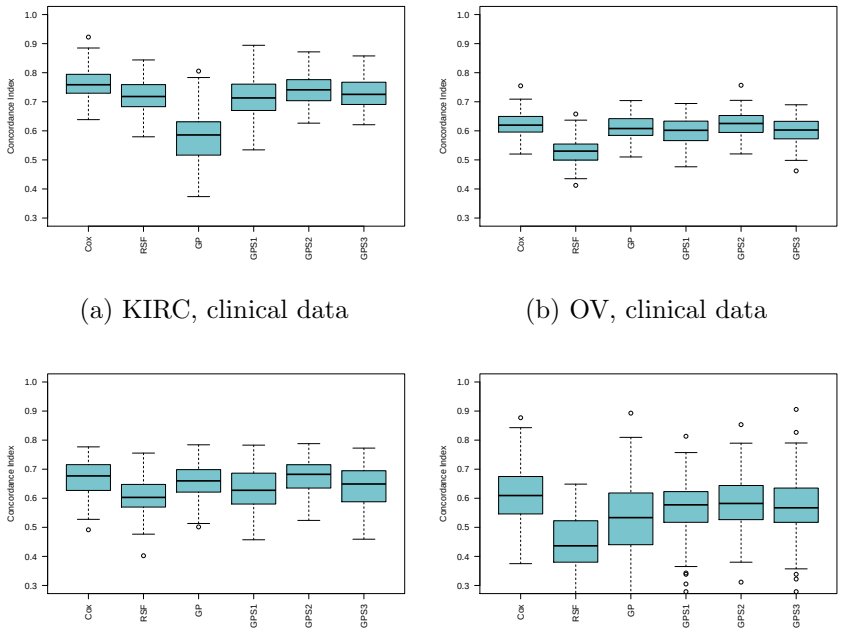
It may be seen that, whilst the Gaussian process model trained on only the uncensored samples fails to train a predictive model, the survival models GPS1, GPS2 and GPS3 have predictive ability comparable with that of the Cox proportional hazards and Random Forest for survival models.

4.3.3 Tothill et al. [156] cancer data

The selection of models were also applied to ovarian cancer clinical and mRNA expression data obtained from Tothill et al. [156] using the R package *curatedOvarianData* [48] (See Methods 4.2.3 for more details). 50 subsets of the data containing 200 training and 50 test samples were run by applying Monte Carlo cross validation, whereby the required number of samples are held out at random, and repeated for the required number of folds. This produced 50 concordance index values per model.

The models were applied to the mRNA data from Tothill et al. [156] using one of two gene sets, OCGS or SRGS. The resulting concordance index values may be found in Figures 4.7a and 4.7c. The models were then also run on data comprising of both mRNA expression data as selected by the gene set and clinical features, and may be found in Figures 4.7b and 4.7d.

When we consider the case of OCGS and no clinical variables in Figure 4.7a, it may be seen that, whereas the concordance index results of the accelerated failure time, Cox proportional hazards and Cox proportional hazards model with elastic-net



(a) KIRC, clinical data

(b) OV, clinical data

(c) GBMF, clinical data

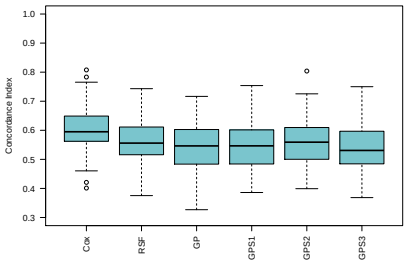
(d) LUSC, clinical data

Figure 4.5: Yuan et al. [171] data, cancers with clinical data

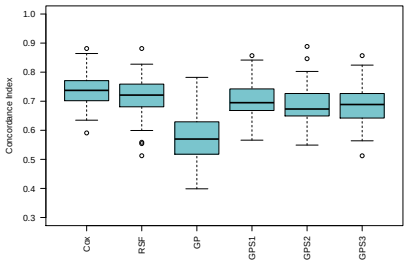
penalisation models are somewhat distributed around 0.5, and the Random Survival Forest and Gaussian process without censored samples models are only marginally better, GPS1, GPS2 and GPS3 outperform all the other models. It is thought that this is a combination of non-linearity and ability to handle high dimensional data.

It was considered whether there is a statistical difference between the results of the GPS models and the models used for comparison in Figure 4.7a. The results of paired Wilcoxon signed rank tests applied to the concordance index values calculated for each model may be found in Table 4.6. The null hypothesis was $\mu_1 = \mu_2$ with a two sided alternative, $\mu_1 \neq \mu_2$. The Holm correction was applied control the familywise error rate.

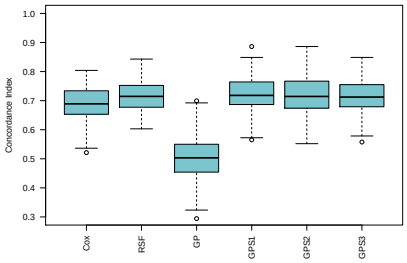
Cohen's d represents the effect size as measured by the difference between the means of two groups. As a rule of thumb, values of d around 0.01 are considered very small, 0.8 is medium, and 2 is considered to be extremely large [137]. Although many of the tests were highly significant ($\alpha = 0.05$), the effect size as measured by the Cohen's d values suggest that that the models fall into two groups, AFT, Cox PH, Coxnet, GP and RSF vs GPS1, GPS2 and GPS3. The values here suggest that there is a large difference between the concordance index values for models in the first group versus the second group, and that there is much less within-group



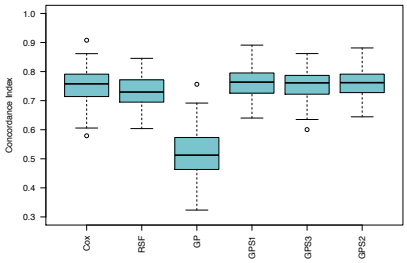
(a) KIRC, SCNA data



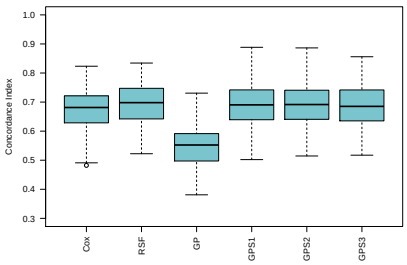
(b) KIRC, SCNA and clinical data



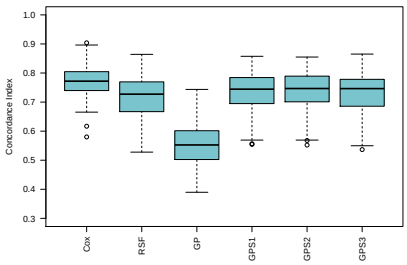
(c) KIRC, mRNA data



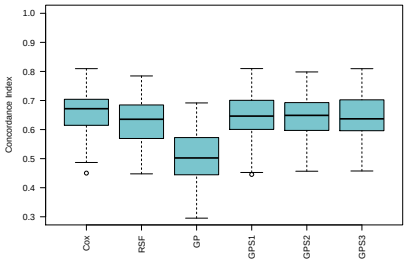
(d) KIRC, mRNA and clinical data



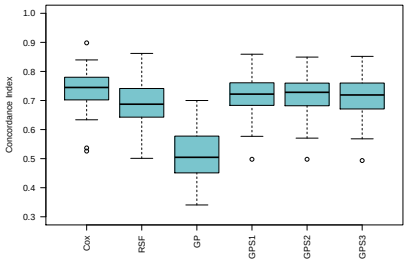
(e) KIRC, miRNA data



(f) KIRC, miRNA and clinical data

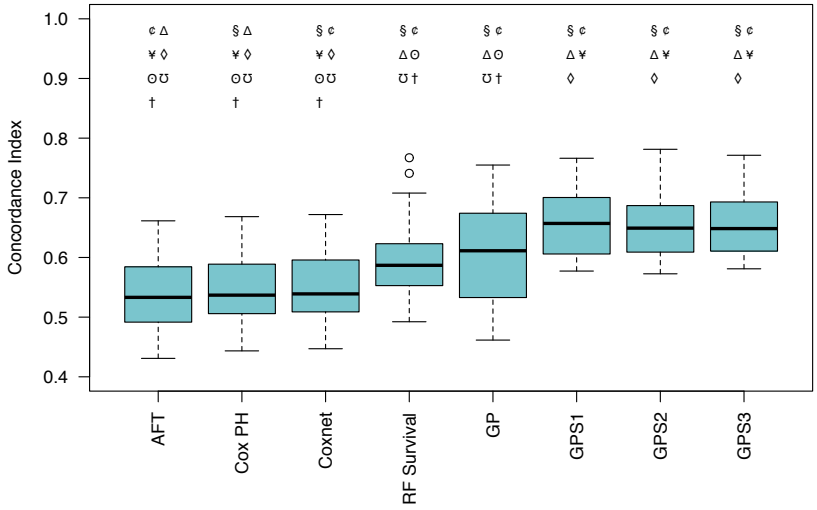


(g) KIRC, proteomic data

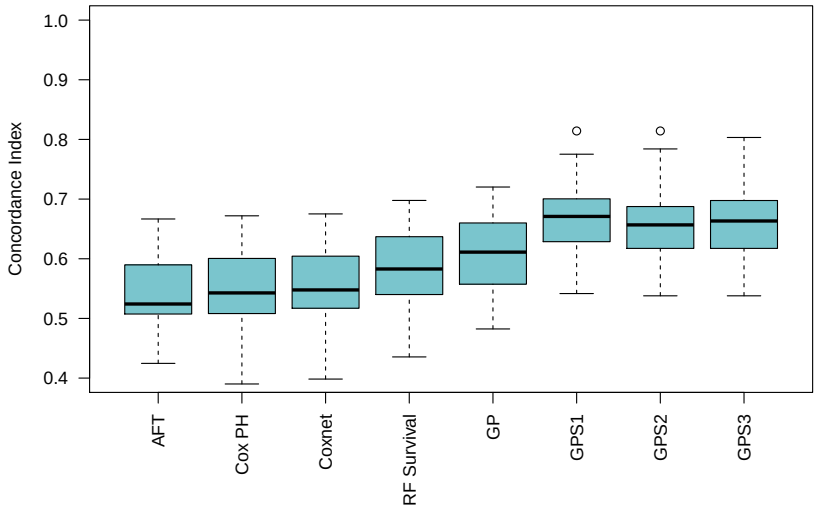


(h) KIRC, proteomic and clinical data

Figure 4.6: Yuan et al. [171] data, KIRC with molecular data

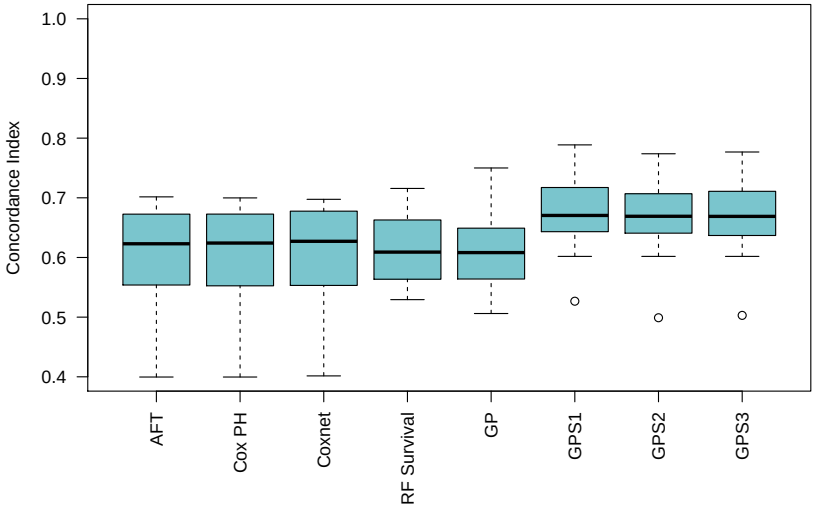


(a) mRNA expression data from gene set OCGS

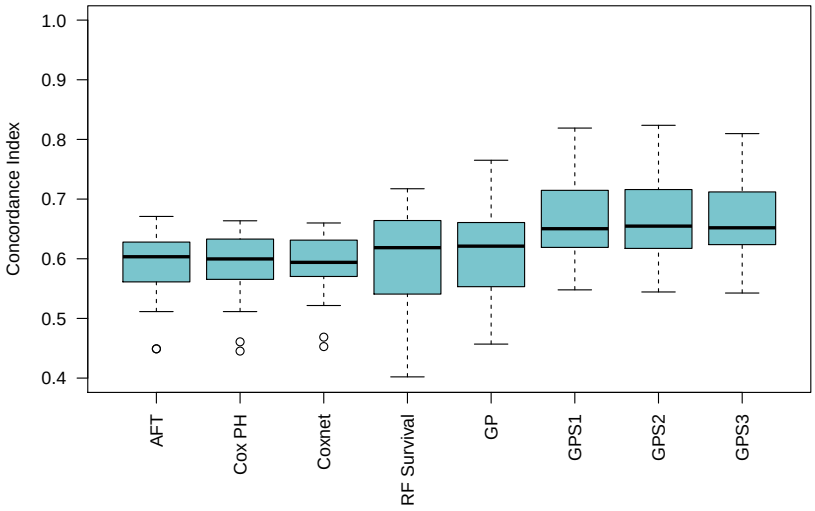


(b) mRNA expression from gene set OCGS and clinical data

Figure 4.7: Tothill et al. [156] data, concordance index of test set predictions. GP, GPS1, GPS2, GPS3, Cox proportional hazards, accelerated failure time, Cox proportional hazards with elastic-net penalisation and Random Survival Forest were applied. a) Symbols show whether a model was significantly different from each other model ($\alpha = 0.05$): § - AFT, ¢ - Coxph, Δ - Coxnet, ¥ - RSF, ◊ - GP, ⊙ - GPS1, ʊ - GPS2, † - GPS3. See Table 4.6 for full details of statistical tests.



(c) mRNA expression data from gene set SRGS



(d) mRNA expression from gene set SRGS and clinical data

Figure 4.7: Tothill et al. [156] data, concordance index of test set predictions. GP, GPS1, GPS2, GPS3, Cox proportional hazards, accelerated failure time, Cox proportional hazards with elastic-net penalisation and Random Survival Forest were applied. a) Symbols show whether a model was significantly different from each other model ($\alpha = 0.05$): § - AFT, ¢ - Coxph, Δ - Coxnet, ¥ - RSF, ◊ - GP, ⊙ - GPS1, ʊ - GPS2, † - GPS3. See Table 4.6 for full details of statistical tests.

variation. GP and RSF may be considered to form an intermediate group, but there is still a large difference in effect size between these models and the GPS models. Within the GPS group, there is little difference between the models, and this was not found to be statistically significant.

Considering Figure 4.7c, in the case of SRGS when clinical features were excluded, all models resulted in similar concordance index values centred around 0.65 with ranges around 0.4–0.75. The median concordance index for the GPS models were marginally higher. When clinical features were included (see Figure 4.7d), there was very little difference in the concordance index values for any models. This suggests that the gene set SRGS is informative, to some extent. However, the large variation in concordance indices achieved between cross-validation folds indicates that the models are not reliably successful, suggesting that this gene set may not be informative for all samples.

4.4 Conclusions

The modelling of right-censored survival data in medicine is a widespread and valuable task. The increasing availability of complex, noisy biomedical data is both a challenge and an opportunity in this regard. New modes of measurement can potentially lead us to new biological and medical insights, providing opportunities for future progress. However, the structure underlying such data types can be more complex than standard methods can fully capture, challenging the abilities of these methods.

This then requires the development of new statistical methods, to use these data to their best advantage. Here this challenge is addressed and the following contributions are made.

- Three models adapting GP regression to right-censored survival data are presented
- It is shown that all three GP models for survival equal or exceed the performance of a range of existing models on both synthetic and a range of real data sets
- This opens up the full toolkit of Gaussian process models to analyse survival data

The Gaussian process modelling framework had already demonstrated its efficacy in a wide range of data modelling contexts, including many biomedical applications. The models presented here extend that framework, providing new ways

Table 4.6: Results of statistical tests applied to Figure 4.7a. Paired Wilcoxon signed rank test, $H_1 : \mu_1 \neq \mu_2$ (i.e. the distribution means of the concordance index values for two models are not equal), Holm multiple testing p-value correction applied. W is the test statistic, CI is the 5–95% confidence interval, and p is the resulting p-value. d is the Cohen's d effect size. p-values in bold are significant at the 5% level.

	Cox PH	Coxnet	RSF	GP	GPS1	GPS2	GPS3
AFT	**p=4.76. $\times 10^{-3}$** (W=3.19, CI=0.00–0.01, d=0.58)	**p=4.58. $\times 10^{-5}$** (W=5.10, CI=0.01–0.02, d=0.93)	**p=1.45. $\times 10^{-3}$** (W=3.70, CI=0.02–0.08, d=0.68)	**p=3.69. $\times 10^{-4}$** (W=4.26, CI=0.03–0.10, d=0.78)	**p=2.36. $\times 10^{-9}$** (W=9.50, CI=0.09–0.14, d=1.73)	**p=2.36. $\times 10^{-9}$** (W=9.42, CI=0.09–0.14, d=1.72)	**p=3.36. $\times 10^{-9}$** (W=9.73, CI=0.09–0.14, d=1.78)
Cox PH		**p=5.65 $\times 10^{-5}$** (W=4.99, CI=0.00–0.01, d=0.91)	**p=4.65 $\times 10^{-3}$** (W=3.22, CI=0.02–0.07, d=0.59)	**p=1.22 $\times 10^{-3}$** (W=3.79, CI=0.03–0.09, d=0.69)	**p=5.86 $\times 10^{-9}$** (W=8.82, CI=0.08–0.13, d=1.61)	**p=6.92 $\times 10^{-9}$** (W=8.68, CI=0.08–0.13, d=1.58)	**p=5.43 $\times 10^{-9}$** (W=8.95, CI=0.08–0.13, d=1.63)
Coxnet			p=9.82. $\times 10^{-3}$ (W=2.83, CI=0.01–0.07, d=0.52)	**p=2.46. $\times 10^{-3}$** (W=3.49, CI=0.02–0.09, d=0.64)	**p=9.57. $\times 10^{-9}$** (W=8.43, CI=0.08–0.13, d=1.54)	**p=1.20. $\times 10^{-8}$** (W=8.29, CI=0.08–0.13, d=1.51)	**p=8.50. $\times 10^{-9}$** (W=8.53, CI=0.08–0.13, d=1.56)
RSF				p=4.13 $\times 10^{-1}$ (W=0.91, CI=-0.02–0.05, d=0.17)	**p=3.36 $\times 10^{-5}$** (W=5.27, CI=0.04–0.09, d=0.96)	**p=8.82 $\times 10^{-5}$** (W=4.80, CI=0.04–0.09, d=0.88)	**p=4.57 $\times 10^{-5}$** (W=5.09, CI=0.04–0.09, d=0.93)
GP					p=5.67 $\times 10^{-3}$ (W=3.07, CI=0.02–0.08, d=0.56)	**p=4.97 $\times 10^{-3}$** (W=3.14, CI=0.02–0.08, d=0.57)	**p=4.97 $\times 10^{-3}$** (W=3.14, CI=0.02–0.08, d=0.57)
GPS1						p=7.03 $\times 10^{-1}$ (W=-0.42, CI=-0.01–0.00, d=0.08)	p=7.03 $\times 10^{-1}$ (W=-0.46, CI=0.00–0.00, d=0.08)
GPS2							p=7.87 $\times 10^{-1}$ (W=0.27, CI=0.00–0.00, d=0.05)

to handle censored survival data, and in the process get more out of the data sets being used.

Chapter 5

Gaussian process feature selection

5.1 Introduction

As cancer research progresses, molecular measurements such as gene expression, miRNA expression and DNA sequencing are producing increasingly high dimensional data sets. These data sets are often have $p \gg n$, where p is the number of features measured and n is the number of samples. For example, of the studies included in the R package *curatedOvarianData*, which was used in Chapter 4, the average number features is 15 046, but 52% of the studies included fewer than 100 samples, and the largest was 578 samples [48]. This 'curse of dimensionality' [17] can cause many different computational and statistical obstacles.

When data sets are high dimensional, it is common for analysis to involve feature selection. This process involves the identification of a subset of features from the whole data set that contain the majority of the information. This subset may then be used to reduce the dimensionality of the problem, or to inform conclusions about underlying processes in the data. Feature selection procedures fall into three main categories [58]: filtering, wrapper or embedded. Filtering procedures involve pre-fitting analysis to identify informative features, and all other features are removed prior to model fitting. Wrapper methods involve the generation of features subsets and fitting using a model as a black box. A score is then used to rate the success of the feature subset. Wrapper procedures are often applied to feature subsets methodically, such as using forward selection or backwards elimination. Common scores include Akaike information criterion (AIC) and Bayesian information criterion (BIC). Embedded methods are those inherent to the model fitting procedure. These

methods run on the whole data set and may produce a measure of feature importance. Examples include Random Forest [20] and elastic net penalisation for generalised linear models [45].

Gaussian processes were selected as the basis for this work due to their flexibility and ability to deal efficiently with high-dimensional data. It is often noted that for Gaussian process regression computational complexity is $\mathcal{O}(n^3)$ due to inversion of the $n \times n$ covariance matrix [14]. However, the dimensionality of the data set also contributes to the runtime complexity. For example, given the squared exponential covariance function, each element is computed as the dot product of two vectors of length p, which has complexity $\mathcal{O}(p)$. For this reason, it is also important that the dimensionality of the input data is not prohibitively large.

The ability to include prior knowledge into statistical machine learning techniques is an important benefit of these methods, particularly in high-dimensional contexts [69]. This may be implemented via the use of priors or by guiding feature selection [85], and may allow, for example, the integration of feature-specific knowledge or the expectation that the number of non-zero parameters is likely to be small [45]. The integration of prior knowledge is expected to encourage a model to favour a biologically relevant regime and hence improve model fitting. In the case of high-dimensional gene expression data, in Chapter 4 the feature set was restricted to lists of genes deemed to be relevant to cancer and resistance to chemotherapy. The identification of a subset of features thought to be biologically relevant allowed the size of the feature space to be reduced, using prior knowledge to select features thought to be informative.

Feature selection in biological contexts is often considered to be useful for understanding relationships and mechanisms of action. For Gaussian processes this is a lesser researched area. Sparse Gaussian process models have been popular, but these are more often applied to reduce the size of the sample space rather than dimensionality [125]. The most basic form of feature selection for Gaussian processes is via the automatic relevance determination (ARD) covariance function. This kernel, which is a generalisation of the squared exponential kernel, has a length hyperparameter per dimension [127]. This allows the length hyperparameters of uninformative features to take very large values, effectively making the relationship of the target with that feature constant. However, in high dimensional contexts, this kernel results in very large numbers of hyperparameters. It is often unlikely that the data set provides enough samples to learn so many hyperparameters adequately, resulting in poor fitting due to lack of information, and resulting in strong reliance on the priors. Simultaneously, the large number of hyperparameters suggests that

the model is susceptible to overfitting, as with the high number of degrees of freedom introduced by large numbers of parameters in statistical models such as linear regression.

In the literature, a selection of other methods have been implemented to provide more complex feature selection for Gaussian processes. Several of these involve placing priors, such as the spike-and-slab prior [49], on parameters in variations of the ARD covariance function [89, 135, 136]. These models have the benefit of being straightforward to implement: a simple change of covariance function and the application of priors. However, they still use very large numbers of hyperparameters in contexts such as gene expression. As a feature selection technique for anisotropic data, Zhou and Suter [172] applied a Fourier transform before clustering features and applying feature selection from a frequency perspective. This process allows the simple squared exponential kernel to be applied, however the results of the feature selection are not easily interpretable. Other studies have used Gaussian processes as a basis for more involved feature selection methods [122, 123, 26]. For example, Pichara and Soto [122] developed a form of local feature selection using many instances of Gaussian process regression to guide feature inclusion. For each feature, a score was calculated at each data point and a Gaussian process model fitted to this score. For each test case, these pre-computed Gaussian processes are then used to determine which features are to be included in the predictive model. This method benefits from the ability for features to be included or discarded flexibly, as may be highly relevant in medical contexts. However, training one model per feature plus one model per test case may become very computationally intensive.

Here two separate feature selection techniques for Gaussian processes have been developed. Firstly, a modification of the ARD covariance kernel has been established to include prior knowledge about relationships in the data. This kernel assigns a length hyperparameter to groups of features, allowing the number of hyperparameters to be kept small. This kernel will be referred to as Informed ARD (IARD) and is detailed in Section 5.2. Secondly, an ensemble method has been developed to allow random subsets of the feature space to be used to generate ensemble predictions for test cases that indirectly involve all features. This model will be referred to as Random Subset Feature Selection (RSFS) and may be found in Section 5.3.

5.2 Informed ARD

For Gaussian processes the most widely accepted embedded feature selection method is the automatic relevance determination (ARD) covariance kernel [127]:

$$k(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \exp\left(-\frac{1}{2}(\mathbf{x}_p - \mathbf{x}_q)^\top M(\mathbf{x}_p - \mathbf{x}_q)\right) \tag{5.1}$$

where $M = \text{diag}(\mathbf{l})^{-2}$.

Here there is a length hyperparameter per feature, which is fitted to be appropriate for each feature. Small values suggest a fast-changing trend, whereas large values represent a flatter, tending towards constant, response. Therefore, features with length hyperparameters tending towards large values are effectively being discounted from the model.

However, in situations where the number of features is large or the number of samples is small, learning values for a length hyperparameter per feature may be problematic. Simultaneously this type of model is prone to overfitting, due to the inclusion of many hyperparameters, and heavy reliance on the mean function, due to not having enough information per feature to adequately learn hyperparameter values.

It was therefore proposed that a version of ARD be developed that was capable of incorporating prior knowledge about the grouping of features. In this way the number of length hyperparameters may be reduced, but the flexibility and feature selection properties of ARD are retained. This method will be referred to as Informed ARD, or IARD.

5.2.1 IARD1

The Informed ARD kernel has a basic structure the same as ARD:

$$k(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \exp\left(-\frac{1}{2}(\mathbf{x}_p - \mathbf{x}_q)^\top M(\mathbf{x}_p - \mathbf{x}_q)\right) \tag{5.2}$$

where $M = \text{diag}(\mathbf{l})^{-2}$.

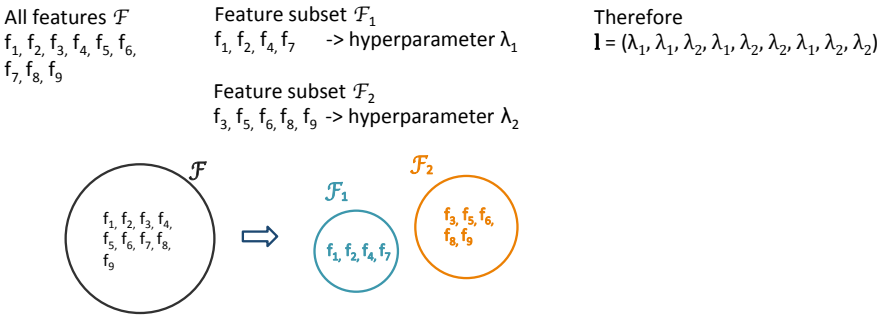
However, the two differ in their assignation of length hyperparameters $\mathbf{l}$ for each feature.

The full set of features, $\mathcal{F}$, is expected to be split into a number of indexed subsets, $\mathcal{F}_1, \ldots, \mathcal{F}_n$, which may be overlapping. These are not necessarily a partition, but must cover $\mathcal{F}$. Each subset of features suggests that the elements of the subset are similar in some way, and hence it is reasonable for them to be assigned the same

length hyperparameter.

Example 1

In the simplest case, of two disjoint subsets covering $\mathcal{F}$, the resulting length hyperparameters may be seen below. Here the two subsets of hyperparameters are assigned one of two different values.



Matters become a little more complicated when the feature subsets are not disjoint. Now, for each feature, contributions must be considered from each subset containing it. Simply, the length hyperparameter for each feature becomes a combination of the length hyperparameters of the subsets to which the feature belongs.

The IARD kernel is constructed as a product of squared exponential kernels, one kernel per feature subset, each with their own length hyperparameter, λ_j. For each feature, each kernel is either on or off, depending on its inclusion in that particular feature set.

$$k_{IARD}(\mathbf{x}_p, \mathbf{x}_q) = \prod_{j=1}^{n} k_j(\mathbf{x}_p, \mathbf{x}_q) \tag{5.3}$$

$$= \sigma_f^{2n} \exp\left(\frac{1}{2}(\mathbf{x}_p - \mathbf{x}_q)^\top \left(\sum_{j=1}^{n} M_j\right)(\mathbf{x}_p - \mathbf{x}_q)\right) \tag{5.4}$$

where there are n feature subsets, $M_j = \lambda_j^{-2} I$, and λ_j is the length hyperparameter for feature subset j, associated with squared exponential kernel j.

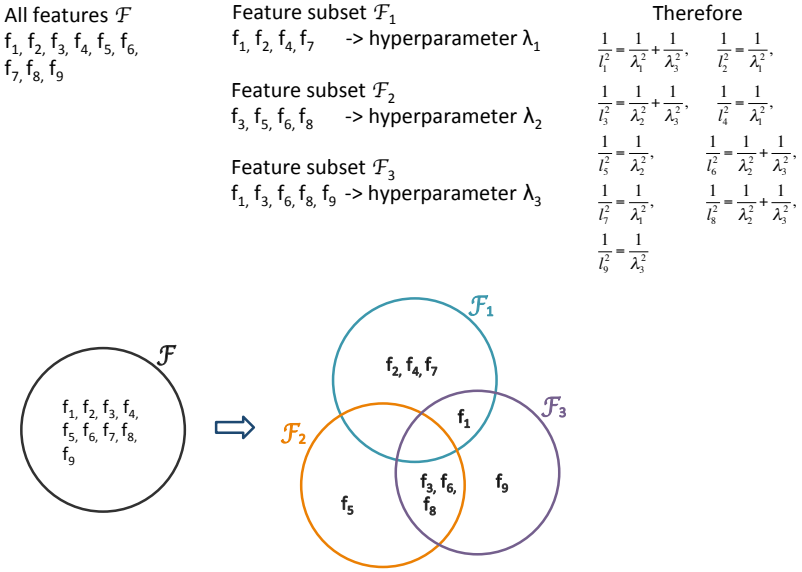
So, given indexed feature subsets $\mathcal{F}_1, \ldots, \mathcal{F}_n \subseteq \mathcal{F}$ with corresponding length hyperparameters $\lambda_1, \ldots, \lambda_n$, the IARD length hyperparameter for feature i is:

$$\frac{1}{l_i^2} = \sum_{j \in \mathcal{J}_i} \frac{1}{\lambda_j^2} \tag{5.5}$$

where $\mathcal{J}_i$ is the index set of subsets of $\mathcal{F}$ containing feature i.

Example 2

In the case of three subsets covering $\mathcal{F}$ but with non-empty intersection, the resulting length hyperparameters must be calculated. For each feature, the length hyperparameters of the subsets to which it belongs are summed to result in the length hyperparameter for that feature.



5.2.2 IARD2

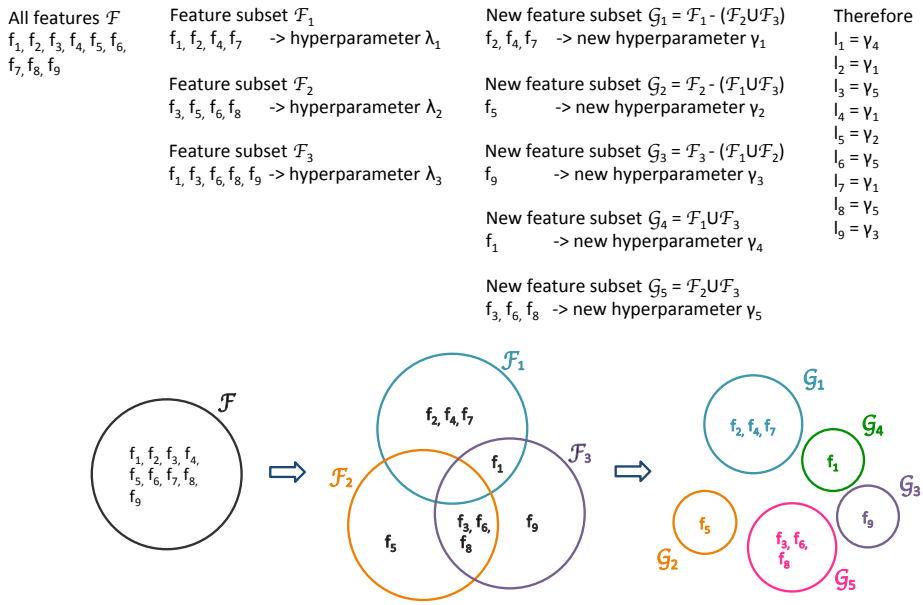
Where feature subsets are not disjoint, the intersection will link more than one, possibly unrelated, subsets together. Consider the case where one subset contains some some informative and some non-informative features, and another contains just some informative features, with the intersection containing a number of informative features. The non-informative features will push the length hyperparameter of the first subset to large values, which will have to be compensated for by the length hyperparameter of the second subset, making it smaller than ideal for the informative features.

In order to address this issue, a pre-processing step has been developed to convert the feature subsets to a partition of the original set of features.

For each combination of intersections between the feature subsets, a new feature subset is created, with a maximum of $\sum_{i=1}^{n} \frac{n!}{i!(n-i)!}$ possible new disjoint subsets. This allows the length hyperparameters of the intersections between feature subsets to vary more flexibly.

Example 3

Here the conversion from covering sets to a partition is applied to three subsets covering $\mathcal{F}$ but with non-empty intersection, as in Example 2. New feature subsets are defined, along with new length hyperparameters, to allow extra flexibility to the intersection between feature sets.



5.3 Random Subset Feature Selection

During feature selection processes, it is often difficult to select a single best model or subset of features, and this approach is often found to give suboptimal results. The use of Bayesian model averaging (BMA) allows this to be avoided, as many possible models are combined. During BMA, many possible models are fitted. For each model, test set predictions are made and the model posterior calculated. For each test set sample, the predictions are then combined to produce ensemble predictions using contributions from all the models. To generate ensemble predictions, predictions are averaged, weighted by the posterior of each model.

Here a method, Random Subset Feature Selection, is investigated whereby BMA is combined with a randomised algorithm for feature subset selection.

Given a data set $\mathcal{D} = (X, y)$, it is expected that of the full set of features $\mathcal{F}$ there is some subset of features $\mathcal{F}_*$ that are informative. In order to identify this informative subset of features, random subsets of features, $\hat{\mathcal{F}}$, are considered and their utility assessed.

It may be considered that, during feature selection, features are designated

as 'on' or 'off'. In a model space containing all possible feature sets, this results in $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ possible models, where n is the full number of features and k is the number of features in the subsets. This number quickly becomes prohibitively large to cover exhaustively, and hence the space must be sampled. Here the sampling is chosen to be random, drawing data subsets from the set of all data subsets with the given number of features without replacement.

Algorithm 5.1: Algorithm to implement Random Subset Feature Selection

Data: Data set $\mathcal{D}$ with features $\mathcal{F}$, model, subsetDimension, nFeatureSubsets
Result: BIC per feature subset $\hat{\mathcal{F}}$

```
begin
    Make a list of all possible subsets, s.t. |F̂| = subsetDimension,
      (# features choose subsetDimension) combinations
    for i ← 1 to nFeatureSubsets do
        Randomly select a feature subset F̂ ⊂ F
        Train model on data subset D̂ containing features F̂
        Calculate BIC and test set predictions
    Compile results from all feature subsets to assess informativeness of
      each feature
    Calculate ensemble predictions, c index etc.
```

Here the form of the BIC used is

$$\mathrm{BIC}_{\mathrm{model}} = -2 \log p(y|X) + d \log(n) \tag{5.6}$$

where d is the number of features in $\hat{\mathcal{D}}$, n is the number of samples, and $\log p(y|X)$ is the posterior output by the model trained on the targets and data, y and X.

This form of the BIC balances the number of features against the fit of the model to the data. In the simple case investigated here the number of features is static, and so the BIC is proportional to the model posterior.

To investigate the informativeness of each feature, the model posterior is marginalised over all models containing the chosen feature. In this case,

$$\begin{aligned} P(f|D) &= \int P(f|M, D) P(M|D) dM \\ &\sim \sum_{\mathcal{F}_i \ni f} \exp(-\mathrm{BIC}_i) \end{aligned} \tag{5.7}$$

where f is the chosen feature of interest and BIC_i is the BIC calculated for feature

set $\mathcal{F}_i$. By comparing the posterior marginalised for each feature, feature importance may be assessed.

For each subset of features, the BIC and test set predictions are calculated. These may be used to generate ensemble predictions, in which BIC is used to create weighted averages of the predictions from each feature subset. $w_i = \exp(-\text{BIC}_i)$ is rescaled such that $\sum_i w_i = 1$, and used as weightings such that for each test sample

$$y_{ensemble} = \sum_{i=1}^{\text{nFeatureSubsets}} w_i \cdot y_i^* \tag{5.8}$$

where y_i^* is the predicted target value for a test sample in feature subset i.

5.4 Methods

5.4.1 Comparison models and abbreviations

The full list of models and abbreviated names for this chapter are listed here. Where relevant, the R package from which the function used is also listed.

- **GP:** GP regression with only the uncensored samples, and censored samples removed
- **GPS3SqExp:** GP for survival data (GPS3) with a squared exponential covariance function
- **GPS3ARD:** GP for survival data (GPS3) with an ARD covariance function
- **GPS3IARD1:** GP for survival data (GPS3) with an IARD1 covariance function
- **GPS3IARD2:** GP for survival data (GPS3) with an IARD2 covariance function
- **GPS3SqExpRSFS:** GP for survival data (GPS3) with a squared exponential covariance function and random subset feature selection
- **GPS3BICForward:** GP for survival data (GPS3) with a squared exponential covariance function with forward selection of features
- **GPS3BICBackward:** GP for survival data (GPS3) with a squared exponential covariance function with backwards elimination of features

- **RF:** Random Forest for regression with censored samples removed, rfsrc (*randomForestSRC*, Ishwaran et al. [73])
- **RSF:** Random Forest for survival, rfsrc (*randomForestSRC*, Ishwaran et al. [73])
- **Coxph:** Cox proportional hazards, coxph (*survival*, Therneau and Grambsch [155])
- **Coxnet:** Cox proportional hazards with elastic-net penalisation, coxph (*survival*, Therneau and Grambsch [155]), glmnet (*glmnet*, Friedman et al. [45], Simon et al. [144])
- **StepCoxph:** Cox proportional hazards with stepwise forward selection of features, coxph (*survival*, Therneau and Grambsch [155]), stepAIC (*MASS*, Venables and Ripley [157])
- **FilterCoxph1:** Cox proportional hazards with univariate filter-based feature selection using Random Forest and repeated cross validation, coxph (*survival*, Therneau and Grambsch [155]), sbf (*caret*, Kuhn [86])
- **FilterCoxph2:** Cox proportional hazards with filter-based feature selection using univariate Cox proportional hazards regression and p-value threshold of 0.05, coxph (*survival*, Therneau and Grambsch [155])

As in Chapter 4, these models are all applied to the same data sets, for all repeats. GPS3 is the Gaussian process for survival model with noise variance correction implemented using the predicted variance of each training target prediction, as detailed in Chapter 3. Where not explicitly stated, RSFS is always applied to GPS3 with the squared exponential covariance function.

Forward selection and backwards elimination procedures involve the use of addition or removal of features from the model. For example, for forward selection, a single feature is added and some measure of the model fit, such as the BIC, is computed. The scores for the addition of each possible variable are compared and the best-performing feature is added to the model. This procedure is continued iteratively until the addition of variables no longer confers improvement in the score, and the resulting feature set is used for the final model. Backwards elimination is performed similarly, but features are removed iteratively from the full feature set, rather than added to an empty set.

In order to compare different feature selection methods, some comparison methods also include feature selection. Explicit feature selection is present in

GPS3BICForward, GPS3BICBackward, Coxnet, StepCoxph, FilterCoxph1 and FilterCoxph2. Due to their mechanisms, implicit feature selection is present in GPS3ARD, GPS3IARD1, GPS3IARD2, GPS3SqExpRSFS, RF and RSF.

5.4.2 Synthetic Data

In order to test the effectiveness of the Informed ARD kernel and the RSFS procedure for feature selection, a series of synthetic data experiments have been carried out. Synthetic data were created using a Gaussian process generatively with zero mean function and ARD covariance function with hyperparameters as detailed in Table 5.1.

The listed proportion of samples were censored randomly to a value drawn from $y_c \sim \mathcal{TN}(\mu = 0, \sigma = 50, 0 < y_c < y)$, ensuring non-informative censoring. Pre-censoring target values were stored to allow comparison with predictions.

The letter following the experiment number indicates which model is being investigated: 'I' for Informed ARD, 'R' for RSFS.

5.4.3 Tothill et al. [156] data

As in Chapter 4, the Tothill et al. [156] data set, accessed via the R *CuratedOvarianData* package [48], was used to test the feature selection procedures on real data. This data set contains survival, clinical and gene expression data from ovarian cancer patients. Here two combinations of molecular and clinical features were used, as seen in Table 5.2. For molecular features, three different options were used: two gene sets, SRGS and OCGS, and adding a random set of 50 genes from the full data set. This applies the filtering feature selection step of including a set of genes thought to be relevant, to reduce the dimensionality of the data set. By including random genes, some data with unknown levels of information are also introduced. All models were run on the same set of features.

5.5 Results

5.5.1 Synthetic Data

Experiment 1IR

It is of interest to compare the predictive ability of the two feature selection methods detailed here. Synthetic data was generated as detailed in Table 5.1. The results of applying the range of models to the synthetic data may be seen in Figure 5.1. Each model was applied to all 50 data repeats and the boxplots show the resulting concordance index values.

Table 5.1: Table of parameters for synthetic data.

Experiment	Dim	# Training Samples	# Test Samples	% Censoring	# Repeats	Hyperparameters σ_n^2	σ_f^2	l
1IR	20	500	100	70	50	0.05	0.7	0.9, 1.1, 1.3, 1.6, 1.65, 1.7, 1.9, 1.95, 2.0, 2.2, 2.6, 2.7, 3.5, 3.7, 4, NA, NA, NA, NA, NA
2R	20	500	100	70	99	0.01	0.5	1.5, 2,3, 2.5, 4,5, 4.5, 5.5, 6, 3, 6.5, 12, 11, 13.5, 10, NA, NA, NA, NA, NA
3R	*	500	100	70	50	0.01	0.5	$\mathcal{G}(k = 2, \theta = 3)$ or NA

* Run for a selection of values, see Section 5.5

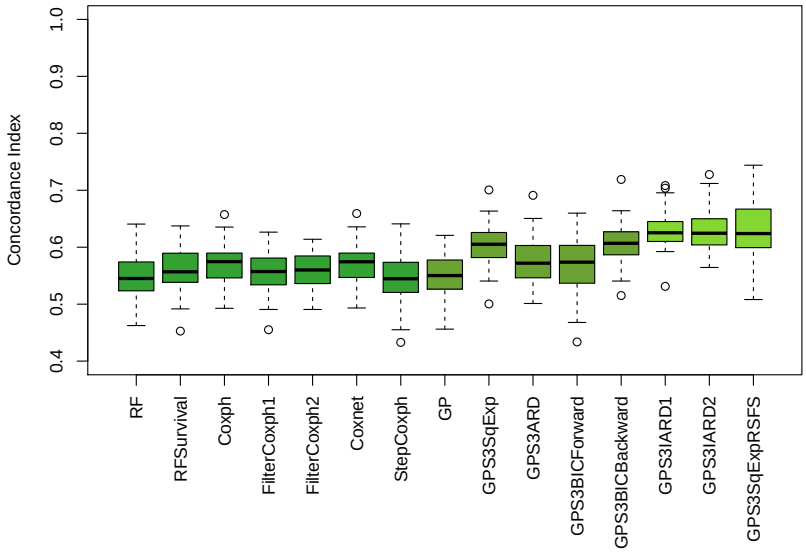


Figure 5.1: Experiment 1IR. Concordance index of test set predictions. Models were applied to the same 50 synthetic data sets, generated using the same hyperparameter values. Boxplots show the median and first and third quartiles, with the whiskers marking 1.5 times the interquartile range from the box.

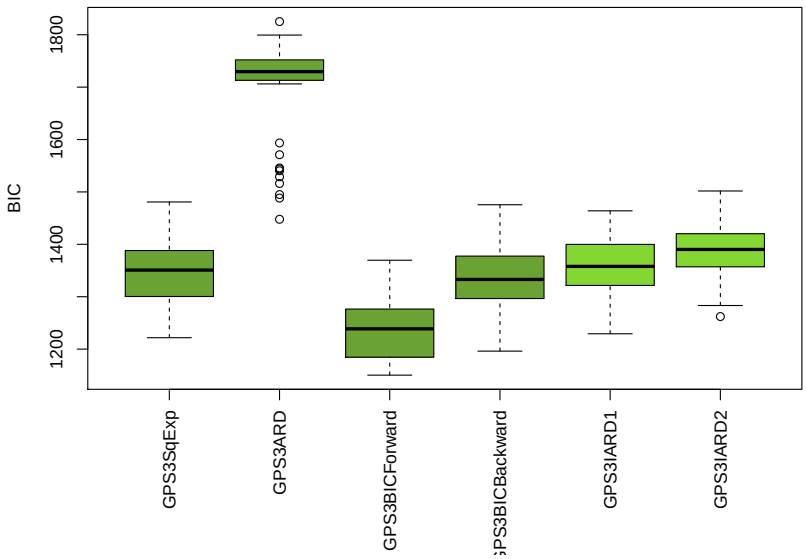


Figure 5.2: Experiment 1IR. The BIC values for each GP model. Lower BIC is more successful. Models were applied to the same 15 synthetic data sets, generated using the same hyperparameter values. Boxplots show the median and first and third quartiles, with the whiskers marking 1.5 times the interquartile range from the box.

Table 5.2: Table of data from Tothill et al. [156] used for each experiment.

Reference	Molecular Features	Clinical Features	Dimension
OCGS+Clin+Rand	OCGS, Random	grade, stage, age	150
SRGS+Clin+Rand	SRGS, Random	grade, stage, age	137

From Figure 5.1 the predictive abilities of the range of models may be compared, using the concordance index values achieved for the test set predictions. One of the best performing models here was GPS with IARD1, with mean concordance index of 0.63 (sd = 0.03). GPS3IARD2 was also successful with mean of 0.63 (sd = 0.04). GPS3SqExpRSFS (mean = 0.63, sd = 0.05) was considered slightly less successful due to larger variation in the range of values.

When considering only the Gaussian process models, other factors may be considered to directly compare the models. Firstly, the trade off between model predictive ability and model complexity may be assessed by considering the BIC. For each GPS model in Figure 5.2, the BIC was calculated as

$$\mathrm{BIC}_{\mathrm{model}} = -2 \log p(y|X) + d \log(n) \tag{5.9}$$

where d is the number of hyperparameters included in the model, n is the number of samples, and $\log p(y|X)$ is the log posterior output by the model trained on the targets y and data X. Recall that the models aim to maximise the log posterior, and hence here a lower BIC is preferable.

Figure 5.2 therefore suggests that the ARD model has not been very successful given the large number of hyperparameters used. When the BIC results are combined with the concordance index results in Figure 5.1, the IARD and RSFS models appear to be a good compromise of high concordance index and low BIC. GPSurvBICForward appears to have achieved a low BIC, but the concordance index values were not particularly high. As this model used the BIC for feature selection scoring, it is therefore likely that this model is overfitted.

Secondly, Figure 5.3 shows the hyperparameter values chosen by each GP model. The boxplots corresponding to each model are grouped and the same colour, with the models ordered as in the plot title and separated by black lines. The number of hyperparameters per model varies, hence so do the number of boxplots. The hyperparameters of the Gaussian process used to generate the data are marked in grey on the plot, and their values may be found in Methods Table 5.1. As the generating GP used a zero mean function and an ARD covariance function, there

are one function variance, one noise variance and 20 length hyperparameters. These generating hyperparameter values may be used as a comparison for the fitted values displayed by the boxplots, as these give an impression of the qualities of the data identified by each model.

The most basic GPS model, GPS3SqExp, used a squared exponential covariance function and hence fitted one function variance, one noise variance and one length hyperparameter. This model may be seen to select length hyperparameter values in the middle of the range of values used by the generating model. However, according to Figure 5.1, the concordance index values achieved by this model were low. It is probable that this model could not fit the complexity of the data using a single length hyperparameter. Similarly to all the GPS models, GPS3SqExp both underestimated the function variance hyperparameter and overestimated the noise variance hyperparameter. This commonly occurred with the GPS models, and is likely due to the additional noise introduced via the censoring. Although the GPS3 model includes a term to account for this, it has not been unusual for GPS models to attribute this additional noise to the noise variance hyperparameter.

The GPS3ARD model fitted 20 length hyperparameter values, and used the same kernel as the GP used to generate the data. However, this model also performed poorly. From Figure 5.3, it may be seen that a large number of runs chose approximately the same values for all the length hyperparameters, and there appears to be even less variation in the values selected for the function and noise variance hyperparameters. It seems likely that the GPS3ARD model failed to learn effectively, and relied heavily on the initial hyperparameter values. This may be due to the large number of hyperparameters, given the number of training samples.

Due to the imposed structure of the generated data, the length hyperparameter groups for IARD1 were four sets: $\mathcal{F}_1 = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F}_2 = \{4, 5, 6, 7, 8, 9, 10, 11, 12\}$, $\mathcal{F}_3 = \{11, 12, 13, 14, 15\}$ and $\mathcal{F}_4 = \{16, 17, 18, 19, 20\}$. For IARD2, these sets became $\mathcal{F}_1 = \{16, 17, 18, 19, 20\}$, $\mathcal{F}_2 = \{11, 12\}$, $\mathcal{F}_3 = \{13, 14, 15\}$, $\mathcal{F}_4 = \{4, 5, 6\}$, $\mathcal{F}_5 = \{7, 8, 9, 10\}$ and $\mathcal{F}_6 = \{1, 2, 3\}$, as the intersections were turned into two additional sets. From Figure 5.3 we notice that both IARD models fit a gene set with a longer length hyperparameter. This represents the feature list containing the non-informative features. By setting the length hyperparameter corresponding to this list at a large value, the models are able to effectively set it as a constant, with very little variation on the scale of the X data.

As the RSFS model is an ensemble model, single-run values are of less interest in this situation. For reference, the RSFS model resulted in chosen length hyperparameters with median 4.3 and first and third quartiles of 2.9 and 6.0 respectively,

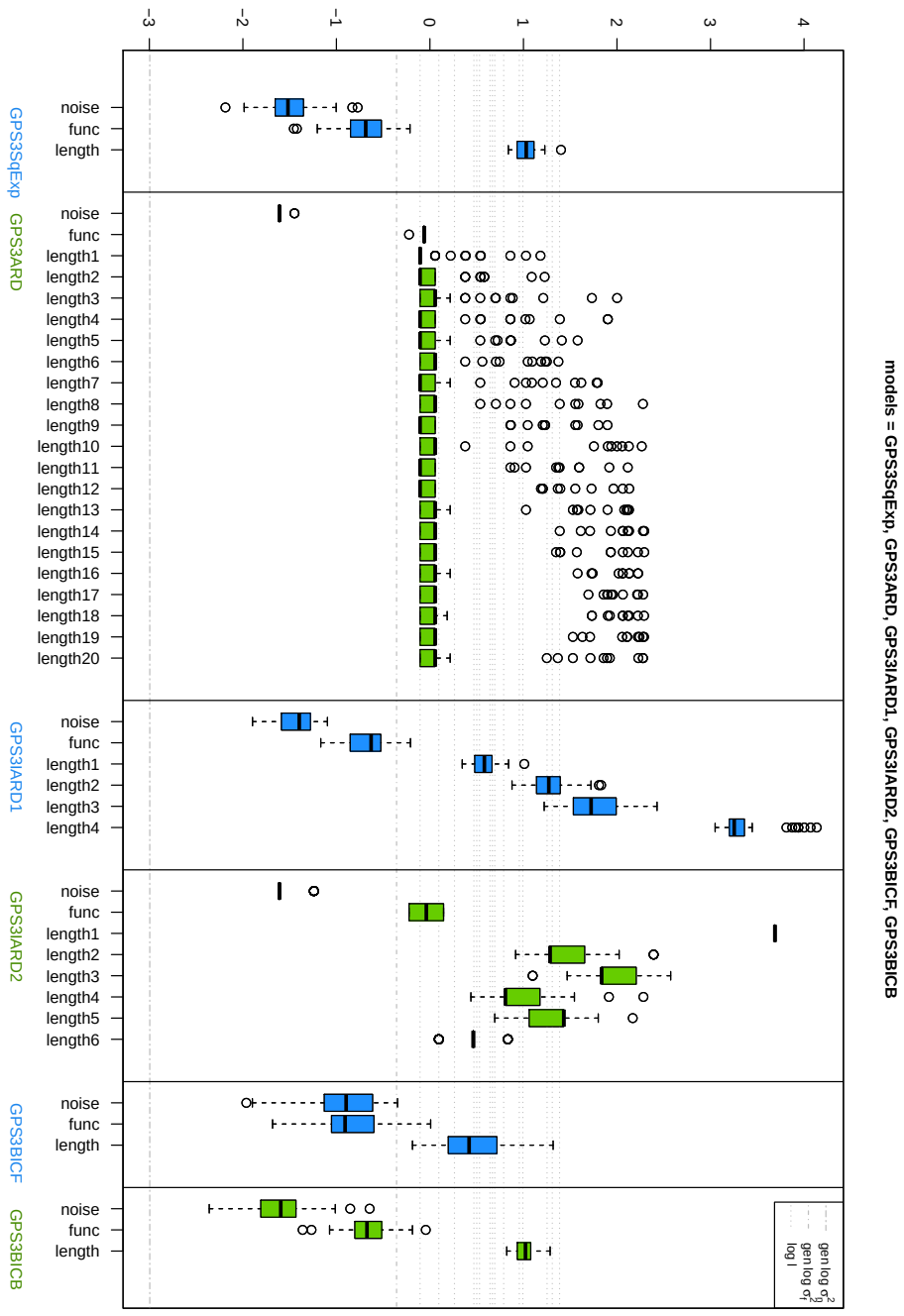


Figure 5.3: Experiment 1IR. Log hyperparameter values chosen by each GP model. Models were applied to the same 50 synthetic data sets, generated using the same hyperparameter values. Boxplots show the median and first and third quartiles, with the whiskers marking 1.5 times the interquartile range from the box. Hyperparameters used to generate data are marked in grey (see Methods Table 5.1 for values).

which are comparable with that of the informative features, x_1 to x_{15}.

A group of models commonly incorporated with feature selection are the Cox proportional hazards models. Here, the basic Cox proportional hazards model does not have feature selection included, but the Coxnet model involves embedded feature selection, StepCoxph applies forward selection of features as a wrapper method, and FilterCoxph1 and FilterCoxph2 both include filtering of features prior to Cox proportional hazards model fitting. It is interesting to note that all of these models achieve similar ranges of concordance index values. When the features retained by these models are considered, those selected by StepCoxph, FilterCoxph1 and FilterCoxph2 vary. FilterCoxph1 and FilterCoxph2 retain features more evenly, and FilterCoxph2 in particular fails to discard features and removes features with longer length scales as often as the non-informative features. StepCoxph placed a strong emphasis on features x_1 to x_8, and x_{12} to x_{20} were retained rarely. Coxnet was found to be less strict, with most features being retained on all runs. These results may be found in Appendix Figure D.1.

Also shown in Appendix Figure D.1, GPSurvBICForward and GPSurvBICBackward resulted in similar trends in retained features, with GPSurvBICForward heavily favouring x_1 to x_7, and rarely retaining features x_{16} to x_{20}. GPSurvBICBackward, as a greedy method, tended to include more features but x_{13} to x_{20} were again retained less often.

Unlike the models with more explicit feature selection, the Random Forest methods do not remove features from the model. However, features are assigned a calculated importance value and these may be used to estimate which features carry the most information. In this case, RF does not appear to differentiate well between features, but RSF assigned features x_1 to x_7 the highest importance. These results may be found in Appendix Figure D.2.

As described in Methods Section 5.3, $\exp(-\text{BIC})$ output by GPS3SqExpRSFS may be marginalised to provide a marginalised probability for each feature. This may be considered analogous to the importance value output by Random Forest models. Appendix Figure D.3 shows that for this experiment the marginalised probabilities for the variables are quite similar, with very little trend. It is therefore likely that many models were given the same weighting and found to be equally important.

Experiment 2R

It was considered what effect the number of unique models has on the predictive ability of the resulting ensemble predictions. For this reason, data was generated as in Table 5.1 and GPS3SqExpRSFS was fitted with differing numbers of feature

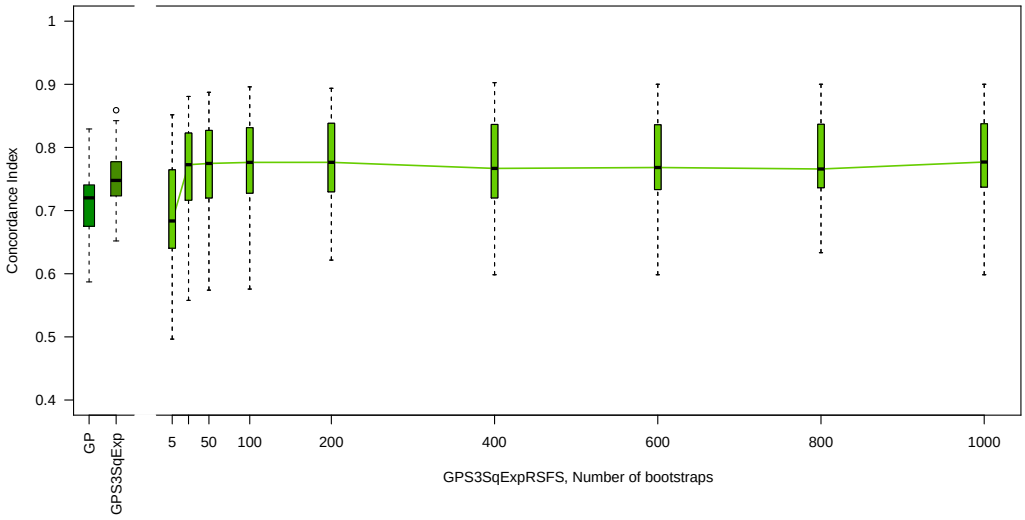


Figure 5.4: Experiment 2R. Boxplots of concordance index values for ensemble predictions as the number of subsets of features are varied. Models are GP, GPS3SqExp, and GPS3SqExpRSFS. Each boxplot represents 99 repeats. GP and GPS3SqExp were applied to each repeat. GPS3SqExpRSFS was applied to each repeat for varying numbers of boostraps.

subsets, where for each subset the GPS3SqExp model was applied to a randomly selected subset of the full set of features. Following the creation of 1000 feature subsets, smaller samples were taken randomly without replacement to provide the other sets of feature subsets. Ensemble predictions were then generated for each set of feature subsets. For comparison, GP and GPS3SqExp were also run on the same data, and all models were applied to each repeat. Figure 5.4 shows the results of 99 repeats of each number of feature subsets.

Statistical tests were carried out to quantify the statistical significance and effect size between models for each number of feature subsets. Two sample Wilcoxon rank sum tests were applied to the concordance index values for each model. The null hypothesis was $\mu_1 = \mu_2$ with a one sided alternative, $\mu_1 < \mu_2$. The Holm correction was applied control the familywise error rate. Cohen's D effect size was calculated similarly, between the concordance index values for the models.

The results may be seen in Table 5.3. The first block compares GP and GPS3SqExp. From the first block of values, it may be seen that the concordance index results for GP versus GPS3SqExp are statistically different ($\alpha = 0.05$), but the effect size as represented by Cohen's D is quite small, suggesting that the two

	GP/GPS3SqExp			
	W statistic	95% CI	P-value	Cohen's D
	7083	0.02–∞	2.0×10^{-6}	0.79

	GPS3SqExp/GPS3SqExpRSFS			
# feature subsets	W statistic	95% CI	P-value	Cohen's D
5	3009	-0.07–∞	1.0×10^{-0}	0.81
25	5898	0.005–∞	4.2×10^{-2}	0.25
50	6151	0.01–∞	1.2×10^{-2}	0.35
100	6248	0.01–∞	8.1×10^{-3}	0.40
200	6242	0.01–∞	8.1×10^{-3}	0.45
400	6148	0.01–∞	1.2×10^{-2}	0.44
600	6415	0.02–∞	2.2×10^{-3}	0.55
800	6521	0.02–∞	9.1×10^{-4}	0.62
1000	6630	0.02–∞	3.4×10^{-4}	0.62

Table 5.3: Table of results statistical tests comparing concordance index values of different models, for different numbers of feature subsets

models do not achieve very different concordance indices.

The second block compares GPS3SqExp to GPS3SqExpRSFS. As seen in Figure 5.4, GPS3SqExpRSFS does not achieve as good concordance index values as the other models when 5 feature subsets are used. However, the results for 25 to 1000 subsets are more interesting. It may be seen that the models are statistically different ($\alpha = 0.05$), though again the effect sizes are still modest. However, they may be seen to rise as the number of feature subsets is increased. This suggests that the GPS3SqExpRSFS model achieved higher predictive ability than GP and GPS3SqExp and, additionally, increasing the number of feature subsets improved the predictive ability, though only slightly.

As the number of feature subsets was increased it was expected that the concordance index of predictions would converge, as a larger number of possible models were sampled. However, it may be seen from Figure 5.4 that this failed to occur. On inspection of the models for each subset, it seems a small number of models are dominating the weightings, which may be seen in Appendix Figure D.4. This is likely due to the inflation caused by $\exp(-\text{BIC})$ for the smallest BIC values. This results in a small number of highly weighted models, and hence the increase in the number of models included is not able to add information to the ensemble predictions. For comparison, the predictions made by GPS3SqExpRSFS were also weighted uniformly and these results may be found in Appendix Figure D.5. There it may be seen that, using uniform weighting, there is a slight improvement in concordance index as the number of feature subsets rises, compared to the $\exp(-\text{BIC})$

weighted version.

It is therefore suggested that the procedures for both the model selection and ensemble generation could be improved. The model selection procedure used here was random selection, and could be improved by some guidance to aid identification of possibly informative models. This method could, for example, blend random and stepwise sampling to further explore models similar to those with low BIC.

Experiment 3R

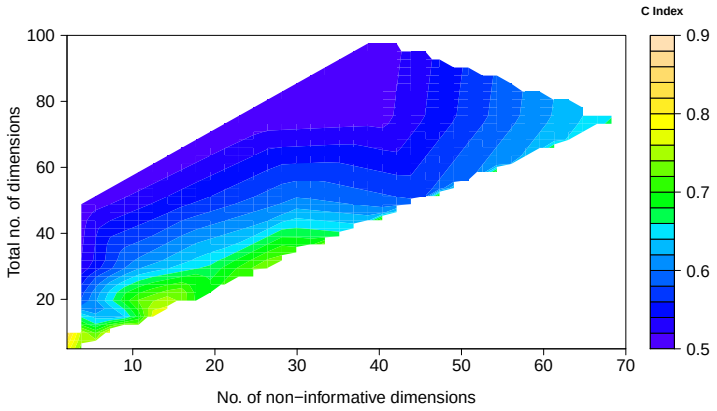
To investigate the Random Subset Feature Selection method more thoroughly, it was proposed to experiment with changing the dimensionality of the data. When generating data, both the dimensionality of the data and the number of non-informative features are set. For this experiment, these are both varied. The length hyperparameters were drawn randomly from $\mathcal{G}(k = 2, \theta = 3)$. As the number of features was increased, the number of feature subsets for the RSFS model was also increased to be approximately 10 times the dimensionality of the data.

For comparison, GP regression with only uncensored samples and GPS3 with squared exponential kernel were also run. Each model was applied to all 50 repeats. Figure 5.5 shows, for each model, an interpolated surface created using the median value from each set of repeats.

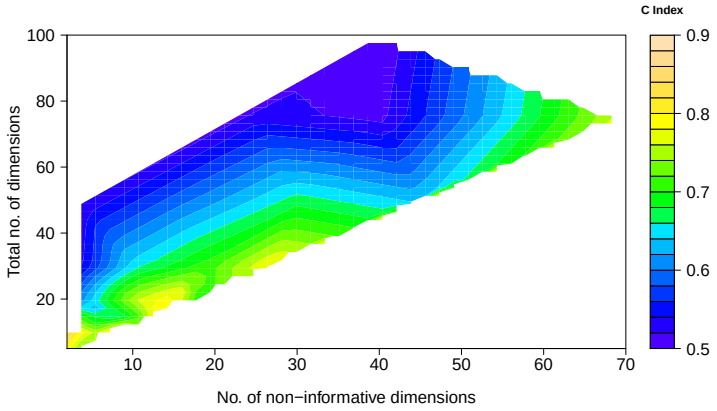
For reference, Appendix Figure D.6 shows the results of the RSFS model, with boxplots showing the concordance index of the model predictions as the total dimensionality and the number of non-informative dimensions changes.

As may be seen in Figure 5.5, the GPS3SqExpRSFS model appears to have higher concordance index values for the higher dimensionality combinations than GP and GPS3SqExp, suggesting that the GPS3SqExpRSFS model is achieving better predictive ability.

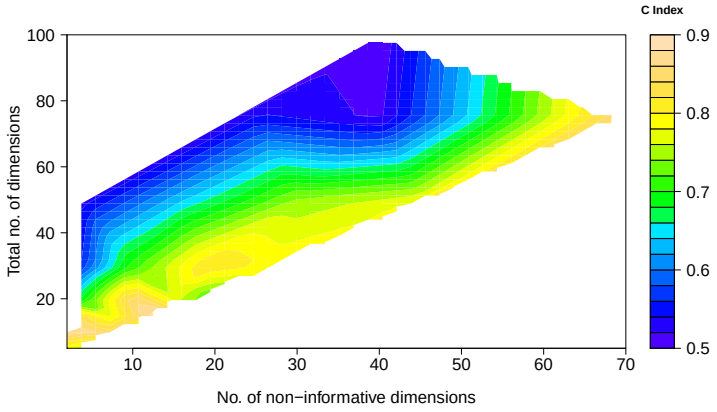
It is interesting to note that for all models, for a set number of features, model predictive ability seems to be better when a greater proportion of features are non-informative. The RSFS feature selection method operates by fitting a GPS model to small subsets of features. In Experiment 2R we saw that the ensemble predictions were dominated by small numbers of highly weighted models. Combined, these two factors suggest that few features are contributing to each feature subset, and also to the ensemble, resulting in low numbers of informative features contributing effectively to the overall model. For data sets with small numbers of informative features, this should have little effect on overall predictive ability, but, for large numbers, the ensemble predictions are unlikely to extract information from enough features to represent the complexity of the data. The predictive ability of the RSFS



(a) GP regression with censored samples removed (GP)



(b) GP for survival data (GPS3SqExp)



(c) GP for survival data with RSFS (GPS3SqExpRSFS)

Figure 5.5: Experiment 3R. Results of running models on synthetic data with changing total number of dimensions (y axis) and number non-informative dimensions (x axis). Interpolated surface created using the median concordance index values from each set of repeats. Colour represents median concordance index, as shown in the scale.

feature selection method therefore is most effective at lower numbers of informative features, given the total number of features in the data set.

It should also be observed that the number of points present on the pre-interpolation grid here is low, and hence any conclusions drawn here should be taken with care. The number of samples appears to be a strong limiting factor in this experiment, and is likely the reason for the lack of predictive ability at high dimensions.

5.5.2 Tothill et al. [156] data

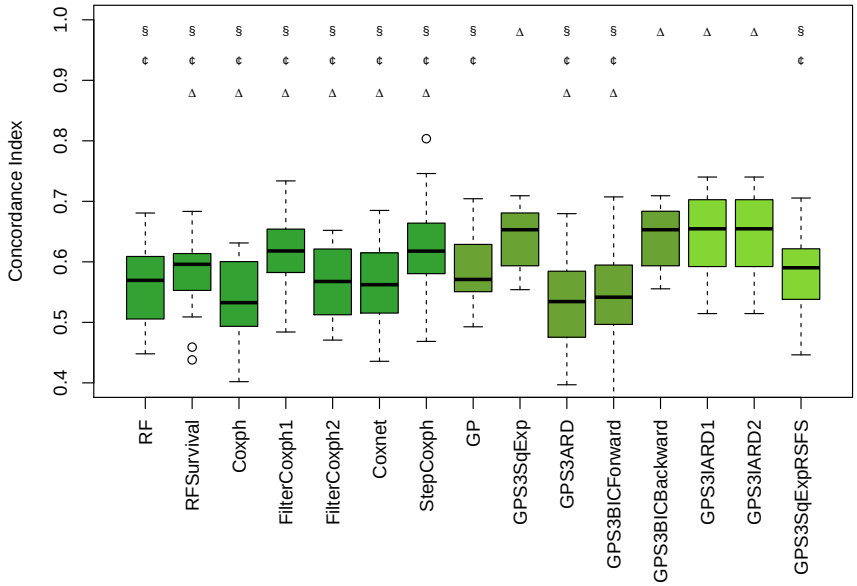
Following application to synthetic data, the same models were applied to two subsets of the Tothill et al. [156] ovarian cancer data set, as detailed in Table 5.2. The results may be found in Figure 5.6.

Overall, the Informed ARD models are achieving equally good concordance index scores as the next best models. In both cases, GPS3 with the ARD kernel achieved much lower scores. It is likely that the data sets used here are too high dimensional for the ARD model to fit well.

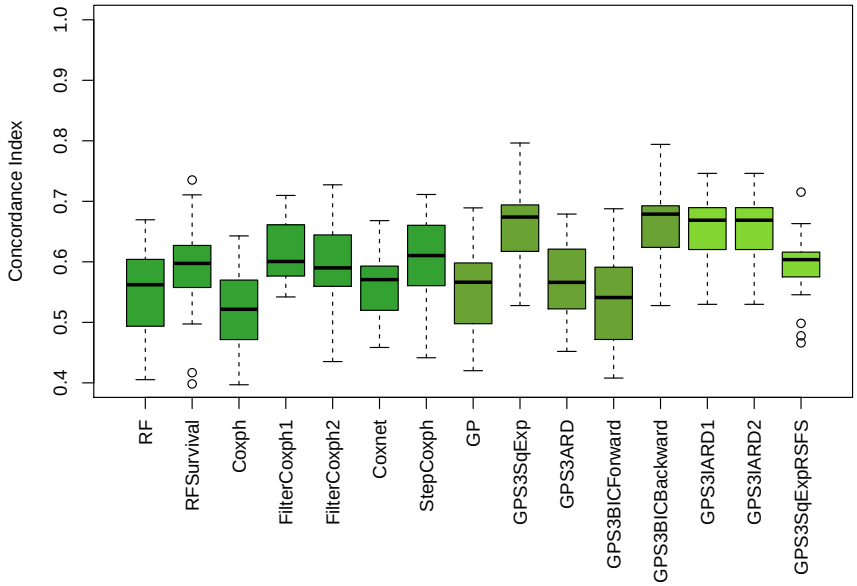
Interestingly, the GPS3SqExpRSFS model does not fare well here. It is suggested that, as in Experiment 2R, the ensemble predictions are being dominated by a small number of models. These models each contain a small number of features, which are unlikely to be capable of particularly good predictions alone due to the nature of expression data. As in Section 5.5.1, further work investigating alternative model selection and ensemble generation procedures is proposed to improve the predictive ability of this model.

Although the difference between models for this data set is not immediately obvious in Figure 5.6, statistical tests were used to assess the differences. Using the results of the 'OCGS+Clin+Rand' data subset in Figure 5.6a, Wilcoxon signed rank tests applied to the concordance index values calculated for each model were applied to compare GPS3IARD1, GPS3IARD2 and GPS3SqExpRSFS to the range of other models. The null hypothesis was $\mu_1 = \mu_2$ with a two sided alternative, $\mu_1 \neq \mu_2$. The Holm correction was applied control the familywise error rate. From Table 5.4 it may be seen that the majority of tests were found to be significant ($\alpha = 0.05$) for all three methods. The effect scores suggest that GPS3IARD1 and GPS3IARD2 performed most similarly to GPS3SqExp and GPS3BICBackward, and GPS3SqExpRSFS performed similarly to GP and RF.

When the same statistical tests are applied to the data in Figure 5.6b, where the SRGS was used, the results are similar, with mostly significant p-values. For the IARD models, Cohen's d effect scores averaged 1.77, whereas GPS3SqExpRSFS



(a) OCGS+Clin+Rand



(b) SRGS+Clin+Rand

Figure 5.6: Results of running models on subsets of the Tothill et al. [156] data set as found in Table 5.2. a) Symbols show whether a model was significantly different from each other model ($\alpha = 0.05$): § - GPS3IARD1, ¢ - GPS3IARD2, Δ - GPS3SqExpRSFS. See Table 5.4 for full details of statistical tests.

averaged lower with 1.16 (not shown).

It should be observed that, given the results of Experiment 3R above, it is likely that the data here are too high-dimensional for the number of samples available. Whilst the common situation of $n << p$ is often the motivation behind feature selection, methods will be limited by the sparsity of samples and lack of information with which to identify useful features and predictive ability suffers as a result.

5.6 Conclusions

As medical testing procedures become increasingly complex, the molecular measurements produced are becoming higher dimensional. Data such as gene expression and DNA sequencing may routinely provide 20 000 dimensional data. When it comes to analysis, this high dimensional data can prove problematic, both in terms of computational power and time.

Feature selection is often used to reduce the dimensionality of a data set and identify informative features. In Chapter 4 filter feature selection methods were applied, both using pre-defined feature sets and univariate Cox proportional hazards models. However, these methods are restrictive and are unlikely to capture complex relationships between features.

For Gaussian processes, the ARD covariance function is capable of implementing embedded feature selection via a change of covariance function. This covariance function assigns a hyperparameter per dimension, which allows uninformative features to effectively be set as constant. However, when applied to high-dimensional data, this method is prone to poor fitting due to the large numbers of hyperparameters used.

In this chapter, two feature selection procedures for Gaussian processes have been proposed and investigated. In order to address the large number of hyperparameters required for the GP ARD kernel, the IARD kernel has been designed to incorporate prior knowledge about relationships between features. Each group of related features are therefore assigned the same length hyperparameter, thereby reducing the total number of hyperparameters. This feature selection procedure has the benefit of being easily integrated, by a simple change of covariance function.

Secondly, an ensemble model utilising Bayesian model averaging using randomly selected feature subsets has also been developed as a wrapper method for Gaussian process models. This model, GPS3SqExpRSFS, allows the space of models to be randomly sampled, and produces ensemble predictions weighted using the BIC.

Table 5.4: Results of statistical tests applied to Figure 5.6a. Paired Wilcoxon signed rank test, $H_1 : \mu_1 \neq \mu_2$ (i.e. the distribution means of the concordance index values for two models are not equal), Holm multiple testing p-value correction applied. W is the test statistic, CI is the 5–95% confidence interval, and p is the resulting p-value. d is the Cohen's d effect size. p-values in bold are significant at the 5% level.

	GPS3IARD1	GPS3IARD2	GPS3SqExpRSFS
RF	**p value = $\mathbf{3.44 \times 10^{-6}}$** (W statistic = 5.90, CI = 0.055–0.113, d = 1.08)	**p value = $\mathbf{3.44 \times 10^{-6}}$** (W statistic = 5.90, CI = 0.055-0.113, d = 1.08)	p value = 5.38×10^{-1} (W statistic = 0.66, CI = -0.021-0.041, d = 0.12)
RFSurvival	**p value = $\mathbf{1.08 \times 10^{-13}}$** (W statistic = 14.09, CI = 0.2–0.267, d = 2.57)	**p value = $\mathbf{1.08 \times 10^{-13}}$** (W statistic = 14.09, CI = 0.2-0.267, d = 2.57)	**p value = $\mathbf{3.03 \times 10^{-11}}$** (W statistic = 10.95, CI = 0.13-0.189, d = 2)
Coxph	**p value = $\mathbf{2.51 \times 10^{-11}}$** (W statistic = 11.05, CI = 0.154–0.223, d = 2.02)	**p value = $\mathbf{2.51 \times 10^{-11}}$** (W statistic = 11.05, CI = 0.154-0.223, d = 2.02)	**p value = $\mathbf{4.06 \times 10^{-6}}$** (W statistic = 5.83, CI = 0.074-0.154, d = 1.06)
FilterCoxph1	**p value = $\mathbf{1.99 \times 10^{-14}}$** (W statistic = 15.46, CI = 0.23–0.3, d = 2.82)	**p value = $\mathbf{1.99 \times 10^{-14}}$** (W statistic = 15.46, CI = 0.23-0.3, d = 2.82)	**p value = $\mathbf{1.09 \times 10^{-10}}$** (W statistic = 10.3, CI = 0.153-0.228, d = 1.88)
FilterCoxph2	**p value = $\mathbf{2.71 \times 10^{-14}}$** (W statistic = 15.15, CI = 0.189–0.249, d = 2.77)	**p value = $\mathbf{2.71 \times 10^{-14}}$** (W statistic = 15.15, CI = 0.189-0.249, d = 2.77)	**p value = $\mathbf{1.18 \times 10^{-8}}$** (W statistic = 8.24, CI = 0.109-0.181, d = 1.5)
Coxnet	**p value = $\mathbf{1.71 \times 10^{-13}}$** (W statistic = 13.78, CI = 0.184–0.248, d = 2.52)	**p value = $\mathbf{1.71 \times 10^{-13}}$** (W statistic = 13.78, CI = 0.184-0.248, d = 2.52)	**p value = $\mathbf{1.06 \times 10^{-7}}$** (W statistic = 7.31, CI = 0.102-0.182, d = 1.34)
StepCoxph	**p value = $\mathbf{1.86 \times 10^{-13}}$** (W statistic = 13.67, CI = 0.232–0.314, d = 2.5)	**p value = $\mathbf{1.86 \times 10^{-13}}$** (W statistic = 13.67, CI = 0.232-0.314, d = 2.5)	**p value = $\mathbf{1.05 \times 10^{-11}}$** (W statistic = 11.50, CI = 0.164-0.234, d = 2.1)
GP	**p value = $\mathbf{3.64 \times 10^{-7}}$** (W statistic = 6.78, CI = 0.048–0.09, d = 1.24)	**p value = $\mathbf{3.64 \times 10^{-7}}$** (W statistic = 6.78, CI = 0.048-0.09, d = 1.24)	p value = 6.85×10^{-1} (W statistic = -0.43, CI = -0.032-0.021, d = 0.08)
GPS3SqExp	p value = 2.22×10^{-1} (W statistic = 1.31, CI = -0.005–0.024, d = 0.24)	p value = 2.22×10^{-1} (W statistic = 1.31, CI = -0.005-0.024, d = 0.24)	**p value = $\mathbf{7.29 \times 10^{-5}}$** (W statistic = -4.75, CI = -0.093–0.037, d = 0.87)
GPS3ARD	**p value = $\mathbf{2.76 \times 10^{-7}}$** (W statistic = 6.90, CI = 0.085–0.157, d = 1.26)	**p value = $\mathbf{2.76 \times 10^{-7}}$** (W statistic = 6.90, CI = 0.085-0.157, d = 1.26)	**p value = $\mathbf{1.62 \times 10^{-2}}$** (W statistic = 2.65, CI = 0.011-0.083, d = 0.48)
GPS3BICForward	**p value = $\mathbf{9.48 \times 10^{-08}}$** (W statistic = 7.36, CI = 0.077–0.137, d = 1.34)	**p value = $\mathbf{9.48 \times 10^{-08}}$** (W statistic = 7.36, CI = 0.077-0.137, d = 1.34)	**p value = $\mathbf{2.67 \times 10^{-2}}$** (W statistic = 2.42, CI = 0.005-0.06, d = 0.44)
GPS3BICBackward	p value = 2.33×10^{-1} (W statistic = 1.27, CI = -0.006–0.024, d = 0.23)	p value = 2.33×10^{-1} (W statistic = 1.27, CI = -0.006-0.024, d = 0.23)	**p value = $\mathbf{5.36 \times 10^{-5}}$** (W statistic = -4.87, CI = -0.092–0.038, d = 0.89)
GPS3IARD1		p value = NaN (W statistic = NaN, CI = NaN-NaN, d = NaN)	**p value = $\mathbf{5.66 \times 10^{-6}}$** (W statistic = 5.69, CI = 0.048-0.101, d = 1.04)
GPS3IARD2	p value = 0 (W statistic = NaN, CI = NaN–NaN, d = NaN)		**p value = $\mathbf{5.66 \times 10^{-6}}$** (W statistic = 5.69, CI = 0.048–0.101, d = 1.04)
GPS3SqExpRSFS	**p value = $\mathbf{5.66 \times 10^{-6}}$** (W statistic = -5.69, CI = -0.101– -0.048, d = 1.04)	**p value = $\mathbf{5.66 \times 10^{-6}}$** (W statistic = -5.69, CI = -0.101– -0.048, d = 1.04)	

This model, unlike the IARD models, does not rely on prior knowledge of the data set.

These feature selection methods have been applied to synthetic and cancer gene expression data and have been found to achieve similar predictive ability to existing methods, such as Coxnet, Random Survival Forest and stepwise feature selection. These methods therefore merit further investigation and development, which will be discussed in Chapter 7.

Chapter 6

REB Array analysis program

Following the development of a qPCR based mutation test for common, actionable mutations in non-small cell lung cancer, melanoma and colorectal cancer by the Cree group, a program for results file analysis has been created. The data generated by this test require analysis prior to the results being clinically accessible, and the process by which this is done is required to be carried out by non-specialists. The analysis program detailed here aims to make data analysis and report preparation simple, fast and reliable, allowing the test to be clinically feasible.

The mutation test was developed as in [81]:
Hugh Kikuchi, Anne Reiman, Jenifer Nyoni, Katherine Lloyd, Richard Savage, Tina Wotherspoon, Lisa Berry, David Snead, and Ian A Cree. Development and validation of a TaqMan array for cancer mutation analysis. *Pathogenesis*, 3 (1):1–8, 2016.

The analysis program was developed by Katherine Lloyd. Ian Cree developed the REB array, Hugh Kikuchi did the testing with assistance from Jenifer Nyoni and Anne Reiman, supervised by Lisa Berry, David Snead, and Tina Wotherspoon.

6.1 Clinical Context

Personalised medicine aims to guide treatment choices using testing and measurements of patient disease state and other factors. By identifying treatments with better predicted outcomes, patient quality of life and survival may be improved, and unnecessary interventions minimised.

Personalised medicine is an important aim in cancer research, originating from the inherent heterogeneity of the disease [97]. Due to the observed abnormal cell responses, cancer is thought to often be triggered by DNA mutations that lead

to disregulation of normal cell mechanisms [159], and hence produce the hallmarks of cancer [62]. Although some research suggests the matter is more complex [34], it is a working assumption in much of cancer research that understanding the underlying mutations within a tumour will allow the development and treatment response of that tumour to be understood. However, techniques to apply this in a clinical setting are currently lacking.

Currently, with the exception of Sanger sequencing, sequencing techniques are little-used in a clinical context, due to their high cost in terms of money, time and expertise. Sanger sequencing is used clinically as a gold-standard test, but cost per sample suggests that this would not be feasible if mutation testing were to be widely implemented, as aimed for by the NHS Personalised Medicine Strategy 2016 [1]. Alternative techniques are therefore required to allow the application of genetic testing in a higher throughput manner. Next generation sequencing techniques, using instruments such as the IonTorrent (Thermo Fisher Scientific) or Illumina systems, are becoming increasingly feasible, but due to cost and expertise required are not yet routinely utilised in practice. During this interim period, it is therefore reasonable to consider more targeted, well established techniques for clinical mutation testing. This is particularly true for triage testing, where a fast, reliable test may be used for the bulk of cases, with next generation sequencing being utilised when required.

One such technique is mutation detection via quantitative polymerase chain reaction (qPCR) gene expression measurement. When testing for mutations, qPCR techniques involve DNA extracted from cancer tissue samples, probes with attached DNA primers, and a polymerase enzyme. The probes consist of a marker, such as a fluorescent reporter, attached to short lengths of DNA that are complementary to the mutation of interest. When these probes are added to the extracted DNA, the complementary sequences adhere and the primer sequences provide a platform for the polymerase enzyme to replicate the DNA sequence. As the name suggests, the reaction then cascades, with replication occurring in an exponential fashion as the number of copies increases. If the mutated gene is present amplification will occur, whereas if the gene does not have the mutation there should be no replication.

The action of the polymerase enzyme requires a series of different temperatures for replication to occur, a sequence of which is referred to as a cycle. These cycles are therefore used as a measure of time. Due to the probe markers, the number of copies of the sequence may be estimated over time. Using a chosen threshold, the cycle number at which the marker passed the threshold is identified, referred to as C_t. For mutation testing, due to the binary nature of the question, a cut-off between mutation and wild type is required. This is usually pre-defined using validation

experiments. If the C_t for a mutation is smaller than the cut-off, the mutation is judged to be present. As a control, a number of wild type probes should also be run, to ensure that the DNA is of sufficient quality and quantity for mutations to be detected if present.

For example, two tests using this technique, the therascreen EGFR RGQ PCR Kit (Qiagen) and the cobas EGFR Mutation Test (Roche Molecular Systems), are currently recommended by NICE [107] but these kits are single-genes and hence, when applied to multiple genes, costs increase.

In clinical contexts, the feasibility of tests is judged by cost in terms of money, time and expertise. Usually, tests are carried out by biomedical or clinical scientists, and time per test is ideally minimised. The complexity of the data analysis following test output being obtained should therefore also kept to a minimum, both to reduce required expertise and to reduce chances for human error. For a test involving gene expression measurement, qPCR is carried out and the resulting amplification measurements are assessed for whether a threshold value was met. This data analysis is well defined and the same process will be applied to all samples. It is therefore a good candidate for automation, whereby the person running the test does not come into contact with the original data further than inputting a file into the analysis program.

The aim of the analysis program would be to take qPCR machine output, apply the analysis using pre-defined rules and thresholds, and output results and reports as required. Details of the experimental methods and data used to develop the test will be outlined in Section 6.2 below, and the subsequent analysis specification in Section 6.3. The technical approach taken and examples may be found in Sections 6.4 and 6.5.

6.2 Data

Kikuchi et al. [81] developed and validated a multi-gene qPCR-based test for actionable mutations relevant to the treatment of colorectal cancer, lung cancer and melanoma. This test aims to maximise information gained whilst simultaneously minimising cost and method complexity.

The genes of interest were EGFR, BRAF, NRAS and KRAS, and the mutations considered have actionable consequences in terms of treatment. NICE recommends the testing of EGFR for all non-small cell lung cancer patients, to assess the utility of prescribing EGFR inhibitors such as gefitinib, erlotinib or osimertinib [108, 106, 111]. For BRAF V600 mutation positive metastatic melanoma,

EGFR	BRAF
Exon 19 del	c.1799T>A
c.2573T>G, c.2572_2573CT>AG	c.1798_1799GT>AA
c.2369C>T	c.1798_1799GT>AG
c.2582T>A	c.1799_1800TG>AT
c.2303G>T	
c.2156G>C	
c.2155G>A	
c.2155G>T	
c.2125G>A	
c.2126A>C	
c.2311_2312insGCGTGGACA	
c.2319_2320ins9	

KRAS	NRAS
c.35G>A	c.37G>C
c.35G>T	c.34G>T
c.38G>A	c.34G>A
c.34G>T	c.38G>A
c.34G>A	c.182A>T, c.181_182CA>TT
c.35G>C	c.35G>A
c.34G>C	c.181C>A
c.436G>A	c.182A>G, c.181_182CA>AG
c.37G>T	c.38G>T
c.183A>C	c.35G>T
c.182A>T	c.183A>T
c.37G>C	c.183A>C
	c.35G>C
	c.37G>T
	c.37G>A
	c.38G>C

Table 6.1: Mutations included on the TaqMan array card, by gene.

NICE recommend a range of BRAF inhibitors including dabrafenib and vemurafenib [110, 109]. It has also been suggested that mutations in KRAS, NRAS, BRAF are predictive of resistance to EGRF inhibitors in colorectal cancer [154]. The mutations tested by Kikuchi et al. [81] for the genes of interest are listed in Table 6.1.

Following DNA extraction from FFPE biopsy tissue, gene expression measurements were carried out using TaqMan microfluidic array cards, which carry out qPCR in 8 sets of 48 wells. Using these cards, the expression of 44 mutations for 7 samples (plus a no-template control) may be measured. In order to test for a mutation, probes corresponding to the mutation of interest are used. If the mutation is present in the loaded DNA the probe binds to it, resulting in a measurable florescence response indicating that the mutation is expressed in the sample. For

each sample, the remaining 4 wells test for a wild type reference per gene, used to check run quality.

As this is a qPCR technique, the presence or absence of a mutation is judged based on when florescence in a well passes a threshold value. The cycle number of this event is referred to as the threshold cycle, or C_t. The threshold value may vary between probes, and is usually set at 3–5 times of the standard deviation of the signal noise above background. The measurement of gene expression using the TaqMan array produces a file containing time course florescence measurements and C_t values for each well.

6.3 Specification

Following the development and validation of the physical methods and materials, for the purposes of the Kikuchi et al. [81] paper, analysis was carried out in Excel (by HK). However, it was decided that this approach was not optimal if the test was to be used in a clinical setting. Instead a dedicated software program was designed for data analysis and report generation.

With input from clinical pathologists (IC and DS), program requirements were determined to be:

- Analysis was required to be carried out in an automated fashion, requiring little user input or tuning
- To conclude that a mutation is present in a sample, the following conditions must be met:
  - The card no-template control must be negative (all wells $C_t > 36.5$)
  - All four sample reference wells must have $C_t \leq 35$
  - The well for the particular mutation must have $C_t < 36.5$
- A report was to be produced for each sample, detailing extraction techniques and mutations present

The subsequent program was designed and written by KL.

6.4 Techniques

R [126] was selected as a suitable programming language for the analysis. Once the analysis was written to specification, the *shiny* [24] package was used to create

an interface with which the user can interact, without requiring knowledge of the language. This package allows the creation of standalone, interactive pages capable of functions such as uploading files, presenting data, plotting and outputting files. The package *knitr* [168] was then used to create PDF reports for each sample containing the requested information.

The program was designed to run as a standalone program in Windows, to open like any other program and to prevent users having to interact with the R console. Running the program starts R and runs the script, and Chrome is started to present the output. When Chrome is closed, R is also closed. The program uses portable versions of R, Chrome and Miktex to enable it to be self-contained and installation to be simply copying over the folder.

In the context of the clinical test under development, gene expression measurements were carried out on a ViiA 7 Real-Time PCR System (Thermo Fisher Scientific). This machine has accompanying software which allows the data produced by runs to be downloaded in a variety of file types. Here .txt was selected as the most suitable, and the analysis was written to read from this file type.

When the program is started, the user is required to locate the appropriate file and upload it to the program. The analysis is then applied automatically. The original file is not altered. When required, the analysis results may be downloaded as a zipped file containing a .pdf report per sample.

The program was written to run on Windows 7 Professional, and has also been tested on Windows XP. The program contains Google Chrome Portable 2.3.0.0, Miktex Portable and R Portable 3.1.2 loaded with the *shiny* [24], *knitr* [168] and *xtable* [33] packages. For generating the PDF reports, the LaTeX packages *amsmath*, *multicol*, *framed* and *geometry* are used.

6.5 Program

When started, the program appears as in Figure 6.1.

As per the instructions, a .txt file produced by the ViiA 7 Real-Time PCR machine is then uploaded. Figure 6.2 shows the Summary tab for one such file. This tab compiles the data from all 7 samples, reporting no-template control status, wild type reference status and any present mutations. If no mutations are detected, mutations are labelled as wild type (WT).

Figure 6.3 shows the contents of a Sample tab. These tabs give a more detailed view of the results for each sample. Time course florescence measurements are plotted for each gene with a line for each mutation. Also present is a small table

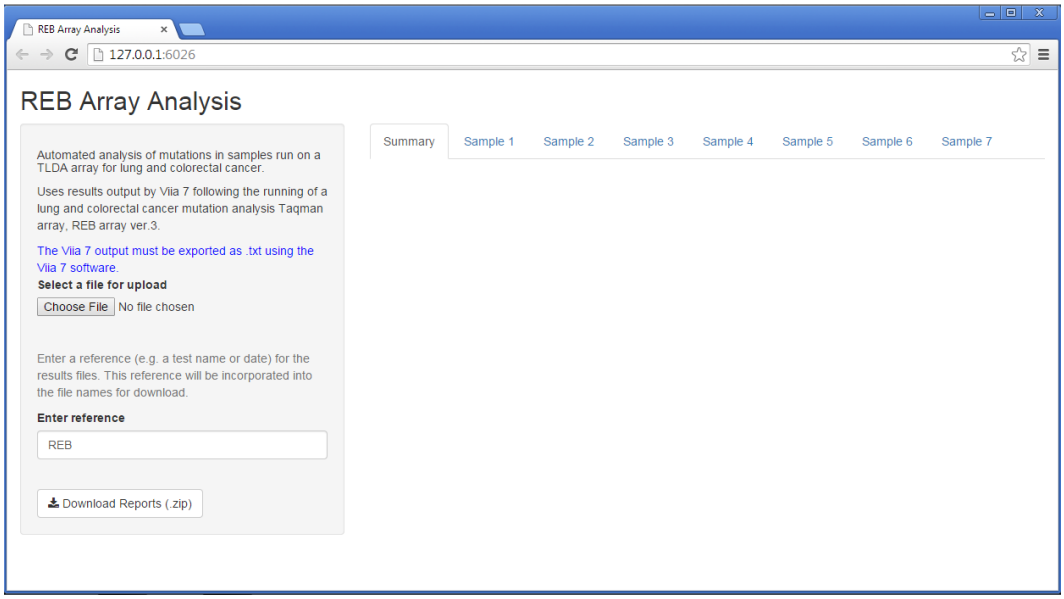


Figure 6.1: Screenshot of program before selecting file

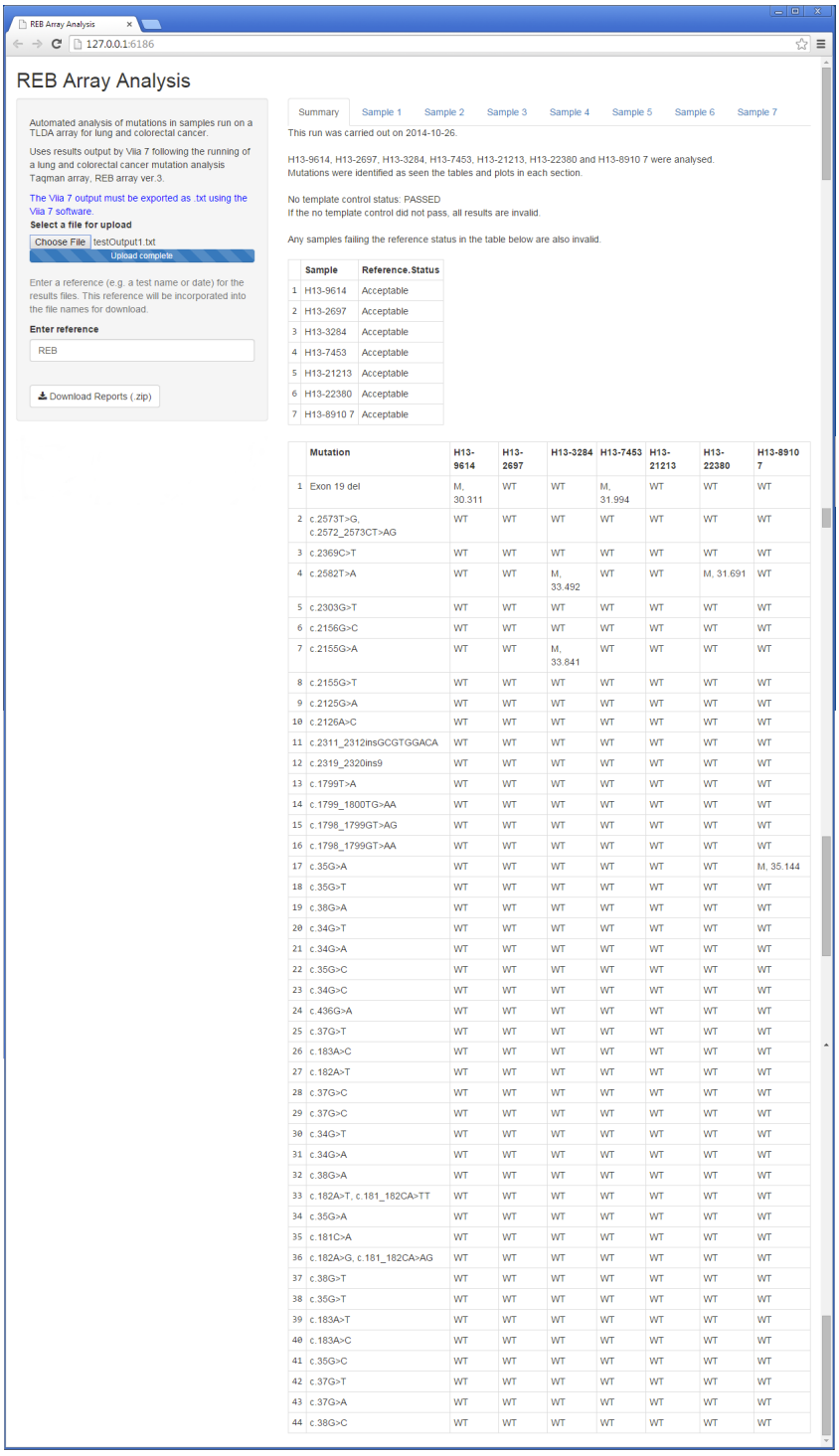
summarising which mutations passed the C_t threshold for each gene.

In addition to the on-screen output, a summary report for each sample is also produced. On clicking the Download Reports button, the reports are generated and then downloaded as a zipped file. Figure 6.4 shows an example report.

6.6 Future generalisations

This approach of developing automated analysis software becomes increasingly important as the complexity of the analysis rises. For example, the movement from qPCR to next generation mutation testing will likely necessitate such software to allow mutation reporting to be feasible within clinical constraints.

Hamblin et al. [60] investigated the appropriateness of setting up a next generation sequencing-based EGFR mutation test for clinical use. As the sequencing platform was the Ion Torrent Personal Genome Machine (Thermo Fisher Scientific), this study utilised the built-in Torrent Suite software for data analysis. This study implemented targeted sequencing using the Ion AmpliSeq Cancer Hotspot Panel (Thermo Fisher Scientific; 46 genes, 189 amplicons), which limits the segments of DNA that are sequenced to those corresponding to the provided primers for amplicon library construction. Here the regions sequenced are 46 genes known to be relevant to cancer. On completion of sequencing, data is passed to the Torrent Suite software, which provides the facility to apply standard or modified analysis



	Sample	Reference.Status
1	H13-9614	Acceptable
2	H13-2697	Acceptable
3	H13-3284	Acceptable
4	H13-7453	Acceptable
5	H13-21213	Acceptable
6	H13-22380	Acceptable
7	H13-8910 7	Acceptable

	Mutation	H13-9614	H13-2697	H13-3284	H13-7453	H13-21213	H13-22380	H13-8910 7
1	Exon 19 del	M, 30.311	WT	WT	M, 31.994	WT	WT	WT
2	c.2573T>G, c.2572_2573CT>AG	WT	WT	WT	WT	WT	WT	WT
3	c.2369C>T	WT	WT	WT	WT	WT	WT	WT
4	c.2582T>A	WT	WT	M, 33.492	WT	WT	M, 31.691	WT
5	c.2303G>T	WT	WT	WT	WT	WT	WT	WT
6	c.2156G>C	WT	WT	WT	WT	WT	WT	WT
7	c.2155G>A	WT	WT	M, 33.841	WT	WT	WT	WT
8	c.2155G>T	WT	WT	WT	WT	WT	WT	WT
9	c.2125G>A	WT	WT	WT	WT	WT	WT	WT
10	c.2126A>C	WT	WT	WT	WT	WT	WT	WT
11	c.2311_2312insGCGTGGACA	WT	WT	WT	WT	WT	WT	WT
12	c.2319_2320ins9	WT	WT	WT	WT	WT	WT	WT
13	c.1799T>A	WT	WT	WT	WT	WT	WT	WT
14	c.1799_1800TG>AA	WT	WT	WT	WT	WT	WT	WT
15	c.1798_1799GT>AG	WT	WT	WT	WT	WT	WT	WT
16	c.1798_1799GT>AA	WT	WT	WT	WT	WT	WT	WT
17	c.35G>A	WT	WT	WT	WT	WT	WT	M, 35.144
18	c.35G>T	WT	WT	WT	WT	WT	WT	WT
19	c.38G>A	WT	WT	WT	WT	WT	WT	WT
20	c.34G>T	WT	WT	WT	WT	WT	WT	WT
21	c.34G>A	WT	WT	WT	WT	WT	WT	WT
22	c.35G>C	WT	WT	WT	WT	WT	WT	WT
23	c.34G>C	WT	WT	WT	WT	WT	WT	WT
24	c.436G>A	WT	WT	WT	WT	WT	WT	WT
25	c.37G>T	WT	WT	WT	WT	WT	WT	WT
26	c.183A>C	WT	WT	WT	WT	WT	WT	WT
27	c.182A>T	WT	WT	WT	WT	WT	WT	WT
28	c.37G>C	WT	WT	WT	WT	WT	WT	WT
29	c.37G>C	WT	WT	WT	WT	WT	WT	WT
30	c.34G>T	WT	WT	WT	WT	WT	WT	WT
31	c.34G>A	WT	WT	WT	WT	WT	WT	WT
32	c.38G>A	WT	WT	WT	WT	WT	WT	WT
33	c.182A>T, c.181_182CA>TT	WT	WT	WT	WT	WT	WT	WT
34	c.35G>A	WT	WT	WT	WT	WT	WT	WT
35	c.181C>A	WT	WT	WT	WT	WT	WT	WT
36	c.182A>G, c.181_182CA>AG	WT	WT	WT	WT	WT	WT	WT
37	c.38G>T	WT	WT	WT	WT	WT	WT	WT
38	c.35G>T	WT	WT	WT	WT	WT	WT	WT
39	c.183A>T	WT	WT	WT	WT	WT	WT	WT
40	c.183A>C	WT	WT	WT	WT	WT	WT	WT
41	c.35G>C	WT	WT	WT	WT	WT	WT	WT
42	c.37G>T	WT	WT	WT	WT	WT	WT	WT
43	c.37G>A	WT	WT	WT	WT	WT	WT	WT
44	c.38G>C	WT	WT	WT	WT	WT	WT	WT

Figure 6.2: Screenshot of Summary tab. Table 1 shows the run status for each sample. Table 2 shows the mutation status for each mutation on the array, for each sample.

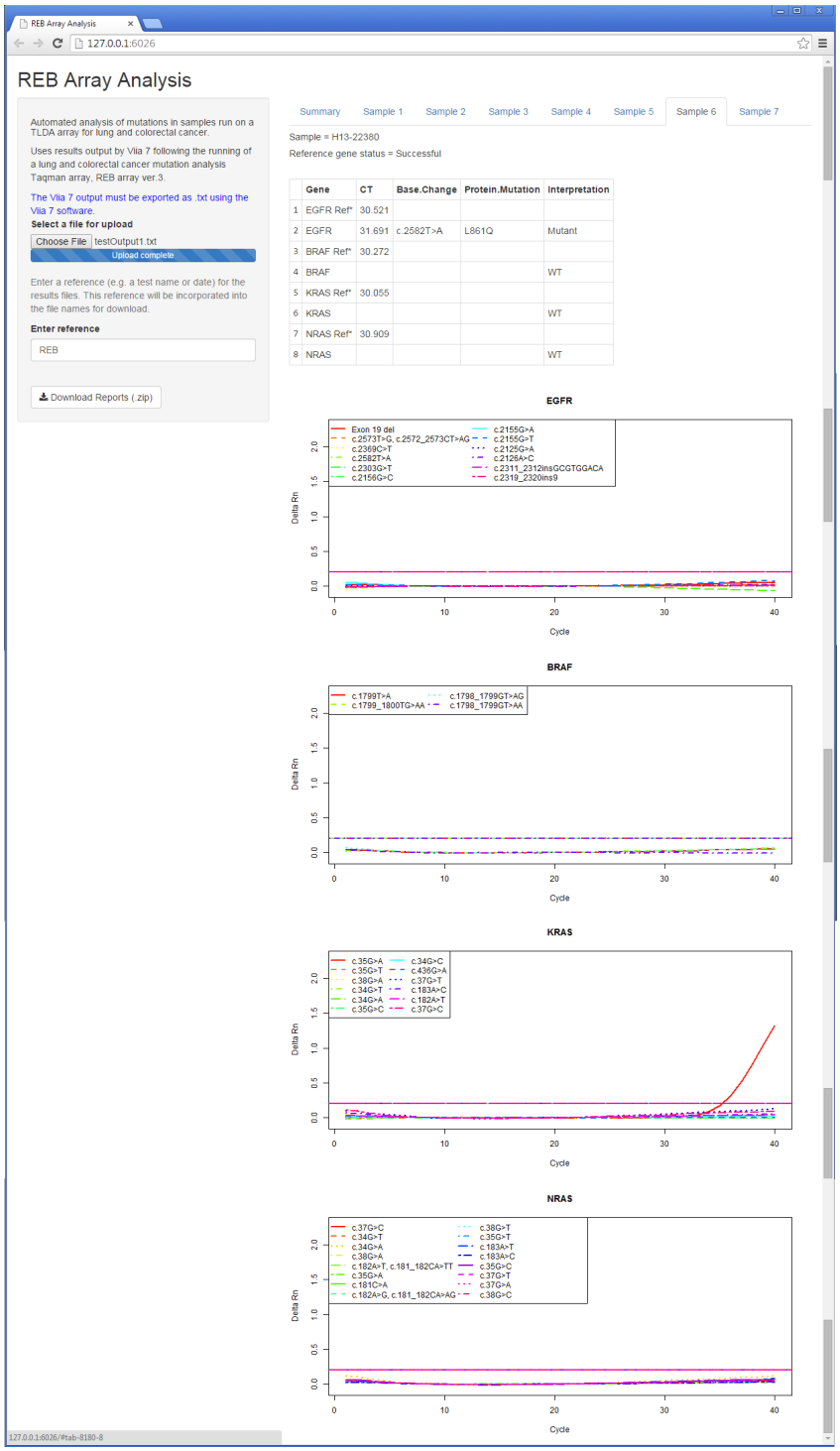


Figure 6.3: Screenshot of Sample 6 tab. Table shows a summary of the mutations present in each gene. Plots show the change in measured fluorescence over time for each gene. Threshold values for each mutation are marked in the same line style as the corresponding time-series.

Supplementary Report

Sample Reference: H13-22380

Run Date: 2014-10-26

Molecular Results

	Gene	CT	Base.Change	Protein.Mutation	Interpretation
1	EGFR Ref*	30.521			
2	EGFR	31.691	c.2582T>A	L861Q	Mutant
3	BRAF Ref*	30.272			
4	BRAF				WT
5	KRAS Ref*	30.055			
6	KRAS				WT
7	NRAS Ref*	30.909			
8	NRAS				WT

* The Ref (reference) result is a control for the wild-type gene.

Genotyping Result

Mutation detected as in the above table.

Testing was carried out using the REB Taqman Array on diagnostic paraffin embedded material selected for testing. This test uses castPCR™, amplifying targeted mutations in exons 18–21 of EGFR; codons 12, 13, 61 and 146 of KRAS; codons 12, 13 and 61 of NRAS; and codon 600 of BRAF. 95% of known mutations are covered. Results are considered reliable down to a concentrate of 2% neoplastic cells. However, at levels as low as this it may be worth considering re-testing of additional material if the result is negative.

Result Interpretation

Mutations detected in EGFR (c.2582T>A; L861Q).

(Pathologist to add note of clinical significance)

1

Figure 6.4: Example report for Sample 6.

workflows. The results of this workflow were lists of the mutations present, sorted into categories: Tiers 1 to 4. Tiers 1 and 2 are known, actionable mutations and known, non-actionable but clinically relevant mutations respectively. Tier 3 mutations are those linked to malignancies, e.g. in My Cancer Genome, COSMIC, ccBioPortal or peer-reviewed literature. Tier 4 mutations have not been previously described. The Tier 1 and 2 mutations are those of interest in a clinical setting. Tier 3 mutations may be of interest if clinical trials are available, and this platform presents the opportunity to test for mutations that are unlikely to be present on qPCR-based tests.

Use of the Torrent Suite software clinically would likely be problematic due to the involved nature of the platform, with the investigation of Tier 3 and Tier 4 mutations requiring the user to access databases. As a simplistic analysis, the inclusion of a .bed file containing a list of mutations of interest would limit the analysis, but other mutations are then lost from the analysis. Clinically, it is important that all present mutations are retained if possible, to allow for treatment research progress.

It is proposed that a second, automated program capable of analysing the results of the Torrent Suite software would greatly improve the clinical applicability of this platform. As the list of mutations of clinical relevance could be pre-defined by a clinician, software capable of extracting these from the list of mutations present in a sample would greatly reduce the specialist knowledge required by the data analyst, and still allow the full list of detected mutations to be retained. As with the qPCR program, clinical reports may then be generated including only the information relevant for the clinician, whilst a full list of mutations is stored elsewhere.

It should also be noted that the list of interesting mutations need not be static. For this type of implementation, it would be highly beneficial for the list to be able to be altered as actionable mutations with targeted therapies, NICE guidelines and available clinical trials change.

For these reasons, software capable of applying well-defined post-Torrent Suite analysis without altering the original data file, applying up to date clinical and NICE guidelines, could be highly beneficial when preparing reports for clinical use.

6.7 Conclusions

With the introduction of personalised medicine the analysis of data following clinical tests, both for single platform and more complex multi-platform combinations, may be expected to greatly increase. For example, the analysis required following next

generation sequencing can be highly in-depth. Similarly, the implementation of a predictive score to aid treatment choices would be likely to involve complex models and statistics. Thus, at the point of contact with the user in a clinical context, it would not be reasonable to expect these analyses to be attempted by hand. It is therefore suggested that an automated approach should be taken, whereby programs taking the output from testing procedures and producing results and reports are utilised.

The benefits of this automated approach are two-fold. Firstly, the specialist knowledge required for the analysis to be carried out is greatly reduced, with little to no interactivity between the method applied and the user. This enables any requirements or standards that need to be met to be integrated into the program, hence ensuring that they are always satisfied. Unlike with the use of more general tools, for which knowledge of applicability and best use are required, an automated analysis tool designed specifically for the data generated by a clinical test could be trusted to implement the pre-decided, appropriate analysis for every sample.

Secondly, the added separation between user and analysis also reduces the chance of user error. The use of read-only files removes the risk of data being overwritten, and the automation of the analysis ensures that all steps of the process are applied. The ability to transfer results automatically to a report without any transcribing by the user again removes an opportunity for user error.

Here a program was developed to provide analysis following mutation testing. Given the well-defined nature of this analysis, with set cut-off values and a pre-determined list of mutations, this was well suited to an automated program, with no user input required. The production of clinical standard reports, containing all required information, allows the program to be used by a biomedical scientist to produce reports suitable for use by clinicians.

It is proposed that this approach, using programs designed to apply highly specific analyses tailored to the data generated, could greatly aid clinical practice as medical tests and models increase in complexity.

Chapter 7

Conclusion

Cancer research has made great strides in improvements to diagnosis, treatment and understanding. However, much of this progress is slow to enter clinical practice or be utilised to highest benefit.

Personalised medicine for cancer is now becoming a feasible aim, with mutation testing and guided treatments starting to be implemented in clinical practice. In order to continue this progress as cancer research becomes increasingly data-rich, tools capable of modelling the high-dimensional data generated as part of molecular clinical tests must therefore be developed, in order to use these data to their best advantage. The ability to apply these tools in a clinical setting would then allow personalised medicine to be used to aid diagnosis and guide treatment decisions.

As seen in Chapter 2, there is currently very little consensus in the literature concerning the prediction of patient response to chemotherapy via gene expression, though many studies have concluded that it is a worthy and feasible goal. The range of laboratory and modelling techniques also vary significantly, and hence so do the data available. Following collation from the studies fulfilling the search criteria, a list of 84 genes identified as predictive by at least two studies was constructed. However, the models used to provide the analysis and predictions were often found to be basic, with strong emphasis on traditional methods such as Cox proportional hazards regression.

Accordingly, three Gaussian process models were developed for survival data in Chapter 3. Gaussian processes were selected due to their suitability for medical data: they are flexible, probabilistic, and model noisy data effectively. Right-censored data may be considered to consist of survival times where for censored samples the survival time is partially missing, and hence the censored times form lower bounds on the true values. The models apply a Gaussian process prior to the set of all

functions relating features to outcome, and infer values for the true values of the censored survival times. In this way, the underlying unmeasured survival times are imputed using information from all available samples, and the resulting model may be used for prediction using unseen samples.

These models were investigated using both synthetic and real data in Chapter 4. Initially the models were tested with right-censored synthetic data created using Gaussian process regression generatively, on which the Gaussian process for survival (GPS) models outperformed the comparison models in terms of predictive ability as measured via concordance index. These results are encouraging, but should be considered in the context of the synthetic data on which they were obtained. The use of Gaussian processes to generate the data means there is an inherent ability for a Gaussian process-based model to fit the data successfully. Additionally, synthetic data form an idealised version of real life data, and are unlikely to include the same levels of complexity and noise present in real data.

Subsequently, the models were applied to molecular data from cancer patients, from two sources. Firstly, the analysis by Yuan et al. [171] of the TCGA data sets for kidney renal clear cell carcinoma, glioblastoma multiforme, ovarian serous cystadenocarcinoma and lung squamous cell carcinoma were recreated, with the addition of the GPS models. This data set consisted of clinical data as well as molecular data from six platforms. The GPS models were found to achieve as good predictive ability as the Cox proportional hazards and Random Survival Forest models, with concordance index values taking similar ranges to these comparison models for all data considered.

Secondly, the models were applied to the gene expression data generated by Tothill et al. [156]. This data set consists of clinical and gene expression measurements, and the GPS models were applied along with a range of comparison models. Here, the list of features was restricted to either of two gene sets, derived from biological pathway knowledge or Chapter 2. Again, the GPS models were found to equal or outperform all other tested models. When applied to gene set OCGS and clinical features, the GPS models were found to achieve statistically significantly ($\alpha = 0.05$) higher concordance index values than the comparison models, with large Cohen's d effect scores.

For the analysis of the Tothill et al. [156] gene expression data, the dimensionality was reduced using the two gene lists. It was therefore proposed that more informative feature selection procedures would be useful. For Gaussian processes this is a lesser researched area, with few easily implemented methods being available. A popular embedded technique for feature selection with Gaussian processes is the

automatic relevance determination (ARD) covariance function. The implementation of this technique is simple, only requiring a change of covariance function, but it can lead to poor fitting due to the large number of hyperparameters required.

In Chapter 5 a modification of the ARD kernel was developed, IARD, to incorporate prior knowledge of similarity between features, such as may be provided by the gene lists used in Chapter 4. The kernel is informed of groups within the features, such as pathways of relevance to the disease, and these are used to reduce the number of hyperparameters required. A second feature selection procedure was also developed whereby a wrapper method was applied to Gaussian processes, to implement Bayesian model averaging with randomly selected feature subsets (RSFS). These two feature selection methods were applied, along with suitable comparison methods, to synthetic data and the Tothill et al. [156] data set. IARD and RSFS were found to achieve similar predictive ability to the existing models and show promise, though I believe their performance could be improved in the future.

RSFS in particular has displayed areas in which further work would be beneficial. Experiments in Chapter 5 suggest that, currently, the method of model averaging when generating ensemble predictions leads to small numbers of highly weighted models dominating the ensemble, with many models with larger BICs contributing much less. To an extent this is expected behaviour; it should be hoped that well-performing models contribute more than lesser-performing models in order to generate good predictions. However, with RSFS in its current form, too few models are being retained to adequately cover more complex feature spaces, and hence predictive ability is suffering. One approach to address this may be the improvement of the feature subset selection procedure. Currently, RSFS uses random selection without replacement for feature subset generation. However, this results in $\frac{n!}{k!(n-k)!}$ possible feature subsets, where n is the total number of features and k is the subset size. For large numbers of features this forms an extremely large model space, which is likely to be sparsely sampled if the number of feature subsets is not increased accordingly. A more involved method for feature subset selection could therefore improve ensemble predictive ability, if feature subsets could be identified that were more likely to provide good predictions. A possible area for investigation is the combination of random and stepwise feature selection, whereby promising areas in model space would be explored more thoroughly before moving on to another randomly selected area. Another alternative for consideration may be the prevention of highly correlated features being present in the same feature subset. As features with high correlation are unlikely to add additional information, the removal of such subsets would allow more informative subsets to be used.

RSFS currently uses feature subsets containing a set number of features, which is determined prior to the model fitting. For simplicity and speed of fitting, in Chapter 5 this was set to reasonably small numbers, but in reality it may be difficult to predict a suitable value. For a given dataset, this value needs to be large enough that inter-feature effects may be captured, but small enough that the underlying GPS model may fit effectively. An interesting extension to RSFS may therefore be to investigate how the feature subset size affects ensemble predictive ability, and explore the widening of the possible feature subset size to multiple values. For example, it may be of use to consider the inclusion of small, medium and large subsets in varying ratios, to attempt to capture dynamics between different numbers of features.

For the IARD methods, the underlying technique is performing reasonably well, as seen when applied to synthetic data. However, for synthetic data, knowledge of the length hyperparameters used when generating the data may be used to accurately inform the selection of feature subsets, resulting in good predictive ability. In reality, this process is quite problematic when being applied to real data. As seen in Chapter 5, the IARD methods did not perform as well, relative to the comparison methods, on real data as on synthetic data, and I believe that a lack of effective feature grouping is at least partially responsible for this effect. When applied to the Tothill et al. [156] data set, the three feature groups were selected to be the features in the chosen gene set (OCGS or SRGS), the randomly selected features, and the clinical features. However, these groups are large and hence the model was fitting many features with the same length hyperparameters. As further work, it would be an interesting progression to investigate the application of biological knowledge when selecting feature groups. The use of databases such as KEGG could be useful in identifying related genes, as well as the use of existing literature and expert knowledge. In this way, the feature grouping inherent to IARD could be used more effectively.

Chapter 6 documents a program developed for the automated analysis of mutation testing data, using previously existing methods for qPCR data analysis and clinically guided parameter values. This program, written with the aim of allowing reproducible, controlled analysis without requiring specialist knowledge, was designed to provide results and reports for clinical use, and is currently used for this purpose at University Hospital of Coventry and Warwickshire. As the complexity of clinical measurement and testing procedures increases, so will the required analysis. Therefore, the automation of such tasks is likely to become increasingly important.

Whilst the analysis carried out by the program in Chapter 6 is simple, similar programs capable of applying well-defined, complex analysis to data without input

from the user would be highly beneficial in the context of personalised medicine. When a model for the prediction of patient response to chemotherapy is developed, for example, this is likely to be complex and requiring specific specialist knowledge. However, for application in a clinical setting to be feasible, this complexity must be separated from the user, to allow reproducible and efficient application of the tool.

In order to build on the progress made using traditional modelling methods, further work is likely to require an emphasis on the flexibility and probabilistic qualities provided by statistical machine learning techniques, to enable the modelling of highly complex interactions. The Gaussian process for survival models developed here have been found to perform well in low-dimensional contexts, and hence may provide a basis for further investigation. However, despite the promise of these models, the extension to higher dimensional data such as that found in molecular cancer data sets would require further development in order to be effective as a clinical tool. The feature selection tools developed here would require further work to meet this aim, but again have shown promise. Additionally, effective use of knowledge of the features likely to be important could greatly benefit the choice and application of relevant models, as the ability to incorporate any prior knowledge is an advantage of machine learning models. Feature selection procedures capable of this, as with IARD, could allow the full volume of information provided by existing cancer research to be put to best advantage.

The actualisation of personalised medicine tools will result in their application in a clinical context. Whilst this is not yet imminent, parallels may be drawn to the increasingly involved analysis required to produce actionable recommendations following tests such as mutation testing. These analyses are required to be carried out by biomedical or clinical scientists by rote - it would be prohibitive for the analysis to require detailed knowledge of the methods being applied or for ad hoc tailoring of the method to be necessary. It is therefore highly preferable for automation of a precisely defined and accepted analysis to be implemented. The advantages of this approach are many, from reduction in opportunities for user error, improvement in reproducibility, to ease of use. When this concept is eventually applied to personalised medicine, the machine learning tools likely to be implemented will similarly benefit from being automated, allowing these highly sophisticated methods to be applied with the same reliability and simplicity as the most basic analyses.

Appendix A

Chapter 2 Supplementary Information

Listing A.1: PubMed search terms

```
(Ovarian Neoplasms[Mesh] OR ((ovarian[tiab] AND cancer[tiab]) NOT medline[sb]) OR ((
    ovarian[tiab] AND carcinoma[tiab]) NOT medline[sb]) OR ((ovarian[tiab] AND
    cancerous[tiab]) NOT medline[sb]) OR ovarian neoplasm*[tw] OR ((ovary[tiab] AND
    cancer[tiab]) NOT medline[sb]) OR ovary neoplasm*[tiab])
AND
("Chemotherapy, Adjuvant"[Mesh] OR "Antineoplastic Agents"[Mesh] OR chemotherapy[
    tiab] OR chemotherap*[tiab] OR Doxorubicin[tiab] OR Cisplatin[tiab] OR
    Carboplatin[tiab] OR taxanes[tiab] OR taxane[tiab] OR Paclitaxel[tiab] OR
    vinorelbine[tiab] OR platinum[tiab] OR Antineoplastic Combined Chemotherapy
    Protocols/therapeutic use[mesh] NOT molecular targeted therapy[mesh])
AND
(Gene Expression/drug effects[Mesh] OR Gene Expression/genetics[Mesh] OR gene
    expression[tiab] OR genetic express*[tiab] OR upregula*[tiab] OR downregula*[tiab]
    OR gene regula*[tiab] OR microarray[tiab] OR microarrays[tiab] OR gene signature*[
    tiab] OR gene expression profiling[mesh] OR microarray analysis[mesh] OR real-time
     polymerase chain reaction[mesh] NOT (MicroRNAs[Mesh] OR microRNA[tiab] OR
    microRNAs[tiab] OR miRNA[tiab] or protein expression[tw]))
AND
(chemoresistance[tw] OR chemo-resistant[tw] OR chemoresistant[tw] OR chemo-
    resistance[tw] OR resistance to chemotherapy[tw] OR (resistance[tiab] AND
    chemotherapy[tiab]) OR (resistant[tiab] AND chemotherapy[tiab]) OR drug resistance
    [tw] OR chemosensitivity[tw] OR chemosensitive[tw] OR (chemoresponse[tiab] AND
    resistance[tiab]) OR (treatment outcome[mesh] AND chemotherapy[tw]) OR gene
    expression regulation,neoplastic[mesh])
AND
```

((Humans[Mesh] NOT cell line, tumor[mesh] NOT cell line[Mesh] NOT Xenograft Model Antitumor Assays[Mesh]) OR (Human[Mesh] AND Tumor cells, cultured[Mesh:noexp] AND (patients[tw] OR patient[tw])) OR ((human[tiab] NOT cell−line[tiab] NOT cell line[tiab] NOT xenograft[tiab]) NOT medline[sb]) OR ((patients[tiab] NOT cell− line[tiab] NOT cell line[tiab]) NOT medline[sb]) NOT (animal model[tiab] NOT medline[sb]))

AND

(Molecular Medicine[mesh] OR molecular diagnostic[tw] OR molecular diagnostic techniques[mesh] OR (resistance[tw] AND predict[tw]) OR (prognosis[tw] AND predict[tw]) OR (outcome[tw] AND predict[tw]) OR (outcome[tw] AND prediction[tw]) OR (prognosis[tw] AND prediction[tw]) OR (relapse[tw] AND prediction[tw]) OR (relapse[tw] AND predict[tw]) OR (recurrence[tw] AND prediction[tw]) OR (recurrence[tw] AND predict[tw]) OR (prognostic[tw] AND prediction[tw]) OR (prognostic[tw] AND predict[tw]) OR (treatment outcome[mesh] AND predict[tw]) OR (response[tw] AND predict[tw]) OR (response[tw] AND prediction[tw]) OR (molecular biologic techniques[tw] NOT medline[sb]) OR (molecular biology techniques[tw] NOT medline[sb]) OR Drug Screening Assays, Antitumor[mesh] OR predictive value of tests [mesh] OR (molecular biological techniques[tw] NOT medline [sb]) OR discriminate[tiab] OR differentiate[tiab] OR categorise[tiab] OR categorize[tiab] NOT imaging[tiab] OR computing methodologies[mesh] OR statistical analysis[tw] OR statistical modelling[tw] OR machine learning[tw] OR supervised[tiab] OR unsupervised[tiab] OR algorithms[mesh] OR multiple linear regression[tiab] OR (regression[tiab] AND analysis[tiab]) OR ((prediction[tiab] OR predictive[tiab] OR predicting[tiab] OR predicts[tiab] OR predict[tiab] OR predictors[tiab] OR predictor[tiab]) AND (development[tiab] OR developed[tiab] OR developing[tiab])) OR predictive value of tests[mesh] OR ((individualised[tiab] OR individualized[tiab] OR personalised[tiab] OR personalized[tiab] OR stratified[tiab]) AND (treatment[tiab] OR medicine[tiab] OR chemotherapy[tiab] OR drug choice[tiab])) OR forecasting[mesh])

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	13
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	13
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	13-16
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	16
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	-
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	16
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	16
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	121-123
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	16-17
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	17
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	17-19
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	19
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	NA
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	NA

Figure A.1: PRISMA Checklist, page 1

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	NA
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	20
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	17-18
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Tables A1-A9
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	126
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Tables A6, 2.4, 2.5
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	NA
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	NA
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	28-32
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	21-39
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	21-39
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	39-42
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	-

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(6): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: www.prisma-statement.org.

Figure A.2: PRISMA Checklist, page 2

QUADAS-2 [164] **results and CEBM 2011 Levels of Evidence** [3]

Study	RISK OF BIAS				APPLICABILITY CONCERNS			Level of Evidence
	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD	FLOW AND TIMING	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD	
Jeong *et. al.* [76]	☺	☺	☺	☺	☺	☺	☺	2
Lisowska *et. al.* [90]	☺	☺	☺	☺	☺	☺	☺	2
Roque *et. al.* [131]	☺	?	☺	☺	☺	☺	☺	2
Li *et. al.* [88]	☺	☺	☺	☺	☺	☺	☺	2
Schwede *et. al.* [141]	☺	☺	☺	☺	☺	☺	☺	2
Verhaak *et. al.* [158]	☺	☺	☺	☺	☺	☺	☺	2
Obermayr *et. al.* [120]	☺	☺	☺	☺	☺	☹	☺	2
Han *et. al.* [61]	☺	☺	☺	☺	☺	☺	☺	2
Hsu *et. al.* [71]	☺	☺	☺	☺	☺	☺	☺	2
Lui *et. al.* [91]	☺	☺	☺	☺	☺	☺	☺	2
Kang *et. al.* [78]	☺	☺	☺	☺	☺	☺	☺	2
Gillet *et. al.* [52]	☺	☺	☺	☺	☺	☺	☺	2
Ferriss *et. al.* [42]	☺	☺	☺	☺	☺	☺	☺	2
Brun *et. al.* [22]	☺	☺	☺	☺	☺	☺	☺	2
Skirnisdottir and Seidal [145]	☺	?	☺	☺	☺	☺	☺	2
Brenne *et. al.* [21]	☺	☺	☺	☺	☺	☺	☺	2
Sabatier *et. al.* [134]	☺	☺	☺	☺	☺	☺	☺	2
Gillet *et. al.* [51]	☺	☺	☺	☺	☺	☺	☺	2
Chao *et. al.* [25]	☹	☺	☹	☺	☺	☺	☹	3
Schlumbrecht *et. al.* [139]	☺	☺	☺	☺	☺	☺	☺	2
Glaysher *et. al.* [54]	☺	☺	☺	☺	☺	☺	☺	2
Yan *et. al.* [169]	☺	☺	☺	☺	☺	☺	☺	2
Yoshihara *et. al.* [170]	☺	☺	☺	☺	☺	☺	☺	2
Williams *et. al.* [165]	☺	☺	☺	☺	☺	☺	☺	2
Denkert *et. al.* [37]	☺	☺	☺	☺	☺	☺	☺	2
Matsumara *et. al.* [98]	☺	☺	☺	☺	☺	☺	☺	2
Crijns *et. al.* [32]	☺	☺	☺	☺	☺	☺	☺	2
Mendiola *et. al.* [103]	?	?	☺	☺	☺	☺	☺	2
Gevaert *et. al.* [50]	?	☺	☺	☺	☺	☺	☺	2
Bachvarov *et. al.* [12]	☹	☺	☺	☺	☺	☺	☺	3
Netinatsunthorn *et. al.* [115]	☺	☺	☺	☺	☺	☺	☺	2
De Smet *et. al.* [35]	☹	☺	☺	☺	☺	☺	☺	3
Helleman *et. al.* [66]	☺	☺	☺	☺	☺	☺	☺	2
Spentzos *et. al.* [148]	☺	☺	☺	☹	☺	☺	☺	2
Jazaeri *et. al.* [74]	☹	☺	☺	☺	☺	☺	☺	3
Raspollini *et. al.* [128]	?	☺	☺	☺	☺	☺	☺	3
Hartmann *et. al.* [63]	☺	☺	☺	☺	☺	☺	☺	2
Spentzos *et. al.* [147]	☺	☺	☺	☺	☺	☺	☺	2
Selvanayagam *et. al.* [142]	?	☺	☺	☺	☺	☺	☺	3
Iba *et. al.* [72]	☺	☺	☺	☺	☺	☺	☺	2
Kamazawa *et. al.* [77]	☺	☹	☺	☺	☺	☺	☺	2
Vogt *et. al.* [161]	☺	☺	☺	☺	☺	☺	☺	2

☺Low Risk ☹High Risk ? Unclear Risk

Table modified from http://www.bris.ac.uk/quadas/resources/

Figure A.3: Bias assessment, including QUADAS-2 and CEBM levels of evidence.

Table A.1: Papers included in systematic review with basic journal and study information. If more than one value is given, the study used multiple different starting gene-sets or found multiple gene signatures.

Study	Journal	No. Samples	No. Genes in Study	No. Genes in Signature
Jeong et al. [76]	Anticancer Res.	487	612	388, 612
Lisowska et al. [90]	Front. Oncol.	127	$> 47\,000$	0
Roque et al. [131]	Clin. Exp. Metastasis	48	1	1
Li et al. [88]	Oncol. Rep.	44	1	1
Schwede et al. [141]	PLoS ONE	663	2632	51
Verhaak et al. [158]	J. Clin. Invest.	1368	11 861	100
Obermayr et al. [120]	Gynecol. Oncol.	255	29 098	12
Han et al. [61]	PLoS ONE	322	12 042	349, 18
Hsu et al. [71]	BMC Genomics	168	12 042	134
Liu et al. [91]	PLoS ONE	737	NS	227
Kang et al. [78]	J. Nat. Cancer Inst.	558	151	23
Gillet et al. [52]	Clin. Cancer Res.	80	356	11
Ferriss et al. [42]	PLos ONE	341	NS	251, 125
Brun et al. [22]	Oncol. Rep.	69	6	0
Skirnisdottir and Seidal [145]	Oncol. Rep.	105	3	2
Brenne et al. [21]	Hum. Pathol.	140	1	1
Sabatier et al. [134]	Br. J. Cancer	401	NS	7
Gillet et al. [51]	Mol. Pharmeceutics	32	350	18, 10, 6
Chao et al. [25]	BMC Med. Genomics	6	8173	NS
Schlumbrecht et al. [139]	Mod. Pathol.	83	7	2
Glaysher et al. [54]	Br. J. Cancer	31	91	10, 4, 3, 5, 5, 11, 6, 6

Table A.1: (continued)

Yan et al. [169]	Cancer Res.	42	2	1
Yoshihara et al. [170]	PLoS ONE	197	18 176	88
Williams et al. [165]	Cancer Res.	242	NS	15 to 95
Denkert et al. [37]	J. Pathol	198	NS	300
Matsumura et al. [98]	Mol. Cancer Res.	157	22 215	250
Crijns et al. [32]	PLoS Medicine	275	15 909	86
Mendiola et al. [103]	PLoS ONE	61	82	34
Gevaert et al. [50]	BMC Cancer	69	$\sim$ 24 000	$\sim$ 3000
Bachvarov et al. [12]	Int. J. Oncol.	42	20 174	155, 43
Netinatsunthorn et al. [115]	BMC Cancer	99	1	1
De Smet et al. [35]	Int. J. Gynecol. Cancer	20	21 372	3000
Helleman et al. [66]	Int. J. Cancer	96	NS	9
Spentzos et al. [148]	J. Clin. Oncol.	60	NS	93
Jazaeri et al. [74]	Clin. Cancer Res.	40	40 033, 7585	85, 178
Raspollini et al. [128]	Int. J. Gynecol. Cancer	52	2	2
Hartmann et al. [63]	Clin. Cancer Res.	79	30 721	14
Spentzos et al. [147]	J. Clin. Oncol.	68	12 625	115
Selvanayagam et al. [142]	Cancer Genet. Cytogenet.	8	10 692	NS
Iba et al. [72]	Cancer Sci.	118	4	1
Kamazawa et al. [77]	Gynecol. Oncol.	27	3	1

Table A.1: (continued)

Vogt et al. [161]	Acta Biochim. Pol.	17	3	0

Table A.2: Papers included in systematic review with tissue information. If more than one value is given, the study used tissue from multiple sources.

Study	Tissue Source	% Cancerous Tissue
Jeong et al. [76]		
Lisowska et al. [90]	Fresh-frozen	NS
Roque et al. [131]	FFPE, Fresh-frozen	min. 70%
Li et al. [88]	FFPE	NS
Schwede et al. [141]		
Verhaak et al. [158]		
Obermayr et al. [120]	Fresh-frozen, Blood	NS
Han et al. [61]		
Hsu et al. [71]		
Liu et al. [91]		
Kang et al. [78]		
Gillet et al. [52]	Fresh-frozen	min. 75%
Ferriss et al. [42]	FFPE	min. 70%
Brun et al. [22]	FFPE	NS
Skirnisdottir and Seidal [145]	FFPE	NS
Brenne et al. [21]	Fresh-frozen effusion, Fresh-frozen	min. 50%
Sabatier et al. [134]	Fresh-frozen	min. 60%
Gillet et al. [51]	Fresh-frozen effusion	NS
Chao et al. [25]		
Schlumbrecht et al. [139]	Fresh-frozen	min. 70%
Glaysher et al. [54]	FFPE, Fresh	min. 80%
Yan et al. [169]	Fresh-frozen	NS
Yoshihara et al. [170]	Fresh-frozen	min. 80%
Williams et al. [165]		
Denkert et al. [37]	Fresh-frozen	NS
Matsumura et al. [98]	Fresh-frozen	NS
Crijns et al. [32]	Fresh-frozen	median = 70%
Mendiola et al. [103]	FFPE	min. 80%

Table A.2: (continued)

Gevaert et al. [50]	Fresh-frozen	NS
Bachvarov et al. [12]	Fresh-frozen	min. 70%
Netinatsunthorn et al. [115]	FFPE	NS
De Smet et al. [35]	Not specified	NS
Helleman et al. [66]	Fresh-frozen	median = 64%
Spentzos et al. [148]	Fresh-frozen	NS
Jazaeri et al. [74]	FFPE, Fresh-frozen	NS
Raspollini et al. [128]	FFPE	NS
Hartmann et al. [63]	Fresh-frozen	min. 70%
Spentzos et al. [147]	Fresh-frozen	NS
Selvanayagam et al. [142]	Fresh-frozen	min. 70%
Iba et al. [72]	FFPE, Fresh-frozen	NS
Kamazawa et al. [77]	FFPE, Fresh-frozen	NS
Vogt et al. [161]	None specified	NS

Table A.3: Papers included in systematic review with histology information. Entries in bold indicate that the study data set was comprised of at least 80% this type.

Study	Sub-type	Stage
Jeong et al. [76]	**Serous**, Endometrioid, Adenocarcinoma	I, II, III, IV
Lisowska et al. [90]	Serous, Endometrioid, Clear cell, Undifferentiated	II, **III**, IV
Roque et al. [131]	Serous, Endometrioid, Clear cell, Undifferentiated, Mixed	IIIC, IV
Li et al. [88]	Serous, Endometrioid, Clear cell, Mucinous, Transitional	II, III, IV
Schwede et al. [141]	Serous, Endometrioid, Clear cell, Mucinous, Adenocarcinoma, OSE	I, II, III, IV
Verhaak et al. [158]	NS	II, III, IV
Obermayr et al. [120]	**Serous**, Non-serous	II, III, IV
Han et al. [61]	**Serous**, Endometrioid, Clear cell, Mucinous, Mixed, Poorly differentiated	II, III, IV
Hsu et al. [71]	NS	**III**, IV

Table A.3: (continued)

Liu et al. [91]	**Serous**	II, **III**, IV
Kang et al. [78]	**Serous**	I, II, III, IV
Gillet et al. [52]	**Serous**	**III**, IV
Ferriss et al. [42]	**Serous**, Clear cell, Other	**III**, IV
Brun et al. [22]	Serous, Endometrioid, Clear cell, Mucinous, Other	III, IV
Skirnisdottir and Seidal [145]	Serous, Endometrioid, Clear cell, Mucinous, Anaplastic	I, II
Brenne et al. [21]	**Serous**, Endometrioid, Clear cell, Undifferentiated, Mixed	II, III, IV
Sabatier et al. [134]	Serous, Endometrioid, Clear cell, Mucinous, Undifferentiated, Mixed	I, II, III, IV
Gillet et al. [51]	**Serous**	III, IV, NS
Chao et al. [25]	NS	NS
Schlumbrecht et al. [139]	**Serous**	III, IV
Glaysher et al. [54]	Serous, Endometrioid, Clear cell, Mucinous, Mixed, Poorly differentiated	IIIC, IV
Yan et al. [169]	Serous, Endometrioid, Clear cell, Mucinous, Transitional	II, **III**, IV
Yoshihara et al. [170]	**Serous**	**III**, IV
Williams et al. [165]	**Serous**, Endometrioid, Undifferentiated	**III**, IV
Denkert et al. [37]	**Serous**, Non-serous, Undifferentiated	I, II, **III**, IV
Matsumura et al. [98]	**Serous**	I, II, III, IV
Crijns et al. [32]	**Serous**	**III**, IV
Mendiola et al. [103]	Serous, Non-serous	**III**, IV
Gevaert et al. [50]	**Serous**, Endometrioid, Mucinous, Mixed	I, III, IV
Bachvarov et al. [12]	**Serous**, Endometrioid, Clear cell	II, III, IV
Netinatsunthorn et al. [115]	**Serous**	**III**, IV
De Smet et al. [35]	Serous, Endometrioid, Mucinous, Mixed	I, III, IV

Table A.3: (continued)

Helleman et al. [66]	Serous, Endometrioid, Clear cell, Mucinous, Mixed, Poorly differentiated	I/II, **III/IV**
Spentzos et al. [148]	**Serous**, Endometrioid, Clear cell, Mixed	I, II, **III**, IV
Jazaeri et al. [74]	**Serous**, Endometrioid, Clear cell, Mixed, Undifferentiated, Carcinoma	II, III, IV
Raspollini et al. [128]	**Serous**	**IIIC**
Hartmann et al. [63]	Serous, Endometrioid, Mixed	II, III, IV
Spentzos et al. [147]	**Serous**, Endometrioid, Clear cell, Mixed	I, II, **III**, IV
Selvanayagam et al. [142]	Serous, Endometrioid, Clear cell, Undifferentiated	**III**, IV
Iba et al. [72]	Serous, Endometrioid, Clear cell, Mixed	I, II, III, IV
Kamazawa et al. [77]	**Serous**, Endometrioid, Clear cell	III, IV
Vogt et al. [161]	NS	NS

Table A.4: Papers included in systematic review with gene expression measurement technique information.

Study	Immunohistochemistry	TaqMan Array	q-RT-PCR	Commercial Microarray	Custom Microarray	RT-PCR
Jeong et al. [76]	✗	✗	✗	✓	✗	✗
Lisowska et al. [90]	✗	✗	✓	✓	✗	✗
Roque et al. [131]	✓	✗	✓	✗	✗	✗
Li et al. [88]	✓	✗	✗	✗	✗	✗
Schwede et al. [141]	✗	✗	✗	✓	✗	✗
Verhaak et al. [158]	✗	✗	✗	✓	✗	✗

Table A.4: (continued)

Obermayr et al. [120]	✗	✗	✓	✓	✗	✗
Han et al. [61]	✗	✗	✗	✓	✗	✗
Hsu et al. [71]	✗	✗	✗	✓	✗	✗
Liu et al. [91]	✗	✗	✗	✓	✗	✗
Kang et al. [78]	✗	✗	✗	✓	✗	✗
Gillet et al. [52]	✗	✓	✗	✗	✗	✗
Ferriss et al. [42]	✗	✗	✗	✗	✓	✗
Brun et al. [22]	✓	✗	✗	✗	✗	✗
Skirnisdottir and Seidal [145]	✓	✗	✗	✗	✗	✗
Brenne et al. [21]	✗	✗	✓	✗	✗	✗
Sabatier et al. [134]	✗	✗	✗	✓	✗	✗
Gillet et al. [51]	✗	✓	✗	✗	✗	✗
Chao et al. [25]	✗	✗	✗	✓	✗	✗
Schlumbrecht et al. [139]	✓	✗	✓	✗	✗	✗
Glaysher et al. [54]	✗	✓	✗	✗	✗	✗
Yan et al. [169]	✓	✗	✗	✗	✗	✗
Yoshihara et al. [170]	✗	✗	✓	✓	✗	✗
Williams et al. [165]	✗	✗	✗	✓	✗	✗

Table A.4: (continued)

Denkert et al. [37]	✗	✗	✗	✓	✗	✗
Matsumura et al. [98]	✓	✗	✓	✓	✗	✗
Crijns et al. [32]	✗	✗	✓	✗	✓	✗
Mendiola et al. [103]	✗	✓	✗	✗	✗	✗
Gevaert et al. [50]	✗	✗	✗	✓	✗	✗
Bachvarov et al. [12]	✗	✗	✓	✓	✗	✗
Netinatsunthorn et al. [115]	✓	✗	✗	✗	✗	✗
De Smet et al. [35]	✗	✗	✗	✗	✓	✗
Helleman et al. [66]	✗	✗	✓	✗	✓	✗
Spentzos et al. [148]	✗	✗	✗	✓	✗	✗
Jazaeri et al. [74]	✓	✗	✗	✗	✓	✗
Raspollini et al. [128]	✓	✗	✗	✗	✗	✗
Hartmann et al. [63]	✗	✗	✗	✗	✓	✗
Spentzos et al. [147]	✗	✗	✗	✓	✗	✗
Selvanayagam et al. [142]	✗	✗	✗	✗	✓	✗
Iba et al. [72]	✓	✗	✓	✗	✗	✗
Kamazawa et al. [77]	✗	✗	✓	✗	✗	✗
Vogt et al. [161]	✗	✗	✗	✗	✗	✓

Table A.5: Papers included in systematic review with basic modelling information. If more than one value is given, the study included patients treated with different treatments.

Study	Patient Prior Chemotherapy Treatment	Model accounts for the different chemotherapies?	Prognostic or predictive?	Model Validated?
Jeong et al. [76]	Platinum-based	✓	Predictive	✓
Lisowska et al. [90]	Platinum/ Cyclophosphamide, Platinum/Taxane	✗	Prognostic	✓
Roque et al. [131]	NS	✗	Prognostic	✗
Li et al. [88]	Platinum/ Cyclophosphamide, Platinum/Taxane	✗	Prognostic	✗
Schwede et al. [141]	NS	✗	Prognostic	✓
Verhaak et al. [158]	NS	✗	Prognostic	✓
Obermayr et al. [120]	Platinum-based	✗	Prognostic	✗
Han et al. [61]	Platinum/ Paclitaxel		Prognostic	✓
Hsu et al. [71]	Platinum/ Paclitaxel + additional treatments	✓	Prognostic	✓
Liu et al. [91]	NS	✗	Prognostic	✓
Kang et al. [78]	Platinum/Taxane		Prognostic	✓
Gillet et al. [52]	Carboplatin/ Paclitaxel		Prognostic	✓
Ferriss et al. [42]	Platinum-based	✓	Predictive	✓
Brun et al. [22]	NS	✗	Prognostic	✗
Skirnisdottir and Seidal [145]	Carboplatin/ Paclitaxel		Prognostic	✗
Brenne et al. [21]	NS	✗	Prognostic	✗
Sabatier et al. [134]	Platinum-based	✗	Prognostic	✓

Table A.5: (continued)

Gillet et al. [51]	NS	✗	Prognostic	✓
Chao et al. [25]	NS	✗	Prognostic	✗
Schlumbrecht et al. [139]	Platinum/Taxane		Prognostic	✗
Glaysher et al. [54]	Platinum, Platinum/Paclitaxel	✓	Predictive	✓
Yan et al. [169]	Platinum-based	✗	Prognostic	✗
Yoshihara et al. [170]	Platinum/Taxane		Prognostic	✓
Williams et al. [165]	NS	✓	Predictive	✓
Denkert et al. [37]	Carboplatin/ Paclitaxel		Prognostic	✓
Matsumura et al. [98]	Platinum-based	✓	Predictive	✓
Crijns et al. [32]	Platinum, Platinum/ Cyclophosphamide, Platinum/ Paclitaxel	✓	Prognostic	✓
Mendiola et al. [103]	Platinum/Taxane		Prognostic	✓
Gevaert et al. [50]	NS	✗	Prognostic	✓
Bachvarov et al. [12]	Carboplatin/ Paclitaxel, Carboplatin/ Cyclophosphamide, Cisplatin/ Paclitaxel	✗	Prognostic	✓
Netinatsunthorn et al. [115]	Platinum/ Cyclophosphamide		Prognostic	✗
De Smet et al. [35]	Platinum/ Cyclophosphamide, Platinum/ Paclitaxel	✗	Prognostic	✓

Table A.5: (continued)

Helleman et al. [66]	Platinum/ Cyclophosphamide, Platinum-based	✗	Prognostic	✓
Spentzos et al. [148]	Platinum/Taxane		Prognostic	✓
Jazaeri et al. [74]	Carboplatin/ Paclitaxel, Cisplatin/ Cyclophosphamide, Carboplatin/ Docetaxel, Carboplatin	✗	Prognostic	✓
Raspollini et al. [128]	Cisplatin/ Cyclophosphamide, Carboplatin/ Cyclophosphamide, Carboplatin/ Paclitaxel	✗	Prognostic	✗
Hartmann et al. [63]	Cisplatin/ Paclitaxel, Carboplatin/Paclitaxel	✗	Prognostic	✓
Spentzos et al. [147]	Platinum/Taxane		Prognostic	✓
Selvanayagam et al. [142]	Cisplatin/ Cyclophosphamide, Carboplatin/ Cyclophosphamide, Cisplatin/ Paclitaxel	✗	Prognostic	✓
Iba et al. [72]	Carboplatin/ Paclitaxel		Prognostic	✗
Kamazawa et al. [77]	Carboplatin/ Paclitaxel		Prognostic	✗
Vogt et al. [161]	Etoposide, Paclitaxel/Epirubicin, Carboplatin/ Paclitaxel	✓	Predictive	✗

Table A.6: Papers included in systematic review with basic modelling information. If more than one value is given, the study used multiple different prediction methods or predicted more than one endpoint.

Study	Prediction	Prediction Method	Predictive Ability
Jeong et al. [76]	Overall Survival	Student's T test, Hierarchical clustering, Compound covariate predictor algorithm, Cox proportional hazards regression, Kaplan-Meier curves, Log-rank test, ROC analysis	'Taxane-based treatment significantly affected OS for patients in the YA subgroup (3 year rate: 74.4% with taxane vs. 37.9% without taxane, p=0.005 by log-rank test)', 'estimated hazard ratio for death after taxane-based treatment in the YA subgroup was 0.5 (95% $CI = 0.31 - -0.82$,$p = 0.005$)'
Lisowska et al. [90]	Chemoresponse, Disease-Free Survival , Overall Survival	Support vector machines, Kaplan-Meier curves, Log-rank test	No genes found to be significant in the training set were significant in the test set, for chemoresponse, DFS or OS

Table A.6: (continued)

Roque et al. [131]	Overall Survival	Kaplan-Meier curves, Log-rank test, Student's T test	'OS was predicted by increased class III β-tubulin staining by both tumor ($HR3.66$, 96% $CI = 1.11$–12.1, $p = 0.03$) and stroma ($HR4.53$, 95% $CI = 1.28$–16.1, $p = 0.02$)'
Li et al. [88]	Chemoresponse (chemoresistant vs. chemosensitive)	Correlation of p-CFL1 staining and chemoresponse	'immunostaining of p-CFL1 was positive in 77.3% of chemosensitive and in 95.9% of the chemoresistant' ($p = 0.014$, $U = 157.5$)
Schwede et al. [141]	Stem cell-like subtype, Disease-Free Survival, Overall Survival	ISIS unsupervised bipartitioning, Diagonal linear discriminant analysis, Gaussian mixture modelling, Kaplan-Meier curves, Log-rank test	OS (p values): Dressman = 0.0354, Crijns = 0.021, Tothill = $4.4E-7$
Verhaak et al. [158]	Poor Prognosis vs. Good Prognosis	Significance analysis of microarrays, Single sample gene set enrichment analysis, Kaplan-Meier curves, Log-rank test	Good or Poor prognosis, likelihood ratio = 44.63

Table A.6: (continued)

Obermayr et al. [120]	Disease-Free Survival, Overall Survival	Kaplan-Meier curves, Cox proportional hazards regression, χ^2 test	'The presence of CTCs six months after completion of the adjuvant chemotherapy indicated relapse within the following six months with 41% sensitivity, and relapse within the entire observation period with 22% sensitivity (85% specificity)'
Han et al. [61]	Complete Response or Progressive Disease	Supervised principal component method	349 gene signature: ROC AUC= 0.702, $p = 0.022$. 18 gene: ROC AUC= 0.614, $p = 0.197$.
Hsu et al. [71]	Progression-Dree Survival	Semi-supervised hierarchical clustering	Good Response vs. Poor Response, $p = 0.021$
Liu et al. [91]	Chemosensitivity, Overall Survival, Progression-Dree Survival	Predictive score using weighted voting algorithm, Kaplan-Meier curves, Log-rank Test, Cox proportional hazards regression	Response of 26 of 35 patients in an independent data set was correctly predicted, patients in the low-scoring group exhibited poorer PFS ($HR = 0.43$, $p = 0.04$), ROC AUC = 0.90(0.86–0.95)

Table A.6: (continued)

Kang et al. [78]	Overall Survival, Progression-Free Survival, Recurrence-Free Survival	Kaplan-Meier curves, Log-rank test, Cox proportional hazards regression, Pearson correlation coefficient	Berchuck dataset: $HR = 0.33$, 95% CI = 0.13–0.86, p = 0.013; Tothill dataset: HR = 0.61, 95% CI = 0.36–0.99, $p = 0.044$
Gillet et al. [52]	Overall Survival, Progression-Free Survival	Supervised principle components method, Cox proportional hazards regression, Kaplan-Meier curves, Log-rank test	'An 11-gene signature whose measured expression significantly improves the power of the covariates to predict poor survival'($p < 0.003$)
Ferriss et al. [42]	Overall Survival	COXEN coefficient, Mann-Whitney U test, ROC analysis, Unsupervised Hierarchical Clustering	Carboplatin: sensitivity = 0.906, specificity = 0.174, PPV = 60%, NPV = 57% (UVA-55 validation set)
Brun et al. [22]	2-year Disease-Free Survival	Student's T test, Principal component analysis, Concordance index, Kaplen-Meier curves, Log-rank test	No genes were found to have prognostic value

Table A.6: (continued)

Skirnisdottir and Seidal [145]	Recurrence, Disease-Free Survival	χ^2 test, Kaplan-Meier curves, Log-rank test, Logistic regression, Cox proportional hazards regression	p53-status ($OR = 4.123$, $p = 0.009$; $HR = 2.447$, $p = 0.019$) was a significant and independent factor for tumor recurrence and DFS.
Brenne et al. [21]	OC or MM, Progression-Free Survival, Overall Survival	Mann-Whitney U test, Kaplan-Meier curves, Log-rank test, Cox proportional hazards regression	Cox Multivariate Analysis: EHF mRNA expression in pre-chemotherapy effusions was an independent predictor of PFS ($p = 0.033$, relative risk = 4.528)
Sabatier et al. [134]	Progression-Free Survival, Overall Survival	Cox proportional hazards regression, Pearson's coefficient correlation score	Favourable vs. Unfavourable: 'sensitivity = 61.6%, specificity = 62.4%, $OR = 2.7$, 95% $CI = 1.7$—4.2; $p = 6.1 \times 10^{-06}$, Fisher's exact test'
Gillet et al. [51]	Overall Survival, Progression-Free Survival, Treatment Response	Linear regression, Hierarchical clustering, Kaplan-Meier curves, Log-rank test	'6 gene signature alone can effectively predict the progression-free survival of women with ovarian serous carcinoma (log-rank $p = 0.002$)'

Table A.6: (continued)

Chao et al. [25]	Chemoresistance	Interaction and expression networks for pathway identification, pathway intersections, betweenness and degree centrality, Student's T test	No statistical measure available. Many genes identified have previously been found experimentally
Schlumbrecht et al. [139]	Overall Survival, Recurrence-Free Survival	Linear regression, Logistic regression, Cox proportional hazards regression, Kaplan-Meier curves, Unsupervised cluster analysis, Log-rank test, Mann-Whitney U test, χ^2 test	'Greater EIG121 expression was associated with shorter time to recurrence (HR = 1.13 (CI = 1.02–1.26), p = 0.021)', 'Increased expression of EIG121 demonstrated a statistically significant association with worse OS (HR = 1.21 (CI1.09–1.35), $p <$ 0.001)'
Glaysher et al. [54]	Chemosensitivity	AIC gene selection, Multiple linear regression	Cisplatin: $R^2_{adj} = 0.836$, $p <$ 0.001
Yan et al. [169]	Chemosensitivity	ANOVA, Student's T test, Mann-Whitney U test	'Immunostaining scores [Annexin A3] are significantly higher in platinum-resistant tumors ($p = 0.035$)'

Table A.6: (continued)

Yoshihara et al. [170]	Progression-Free Survival	Cox proportional hazards regression, Ridge regression, Prognostic index, ROC analysis, Kaplan-Meier curves, Log-rank test	'Prognostic index was an independent prognostic factor for PFS time ($HR = 1.64$, $p = 0.0001$)', sensitivity = 64.4%, specificity = 69.2%
Williams et al. [165]	Overall Survival	COXEN score, Kaplan-Meier curves, Student's T test, ROC analysis, Spearman's rank correlation coefficient, Logistic regression, Log-rank test	Carboplatin and Taxol: sensitivity = 77%, specificity = 56%, $PPV = 71\%$, $NPV = 78\%$
Denkert et al. [37]	Overall Survival	Semi-supervised analysis via Cox scoring, Principal components analysis, Kaplan-Meier curves, Log-rank test, Cox proportional hazards regression	Duke et al.: 'clinical outcome is significantly different depending on the OPI ($p = 0.021$), with an HR of 1.7 (CI 1.1-–2.6)'
Matsumura et al. [98]	Taxane sensitivity, Overall Survival	Hierarchical clustering, Kaplan-Meier curves, Log-rank test	'Patients in the YY1-High cluster who were treated with paclitaxel showed improved survival compared with the other groups ($p = 0.010$)'

Table A.6: (continued)

Crijns et al. [32]	Overall Survival	Supervised principal components method, Cox proportional hazards regression, Kaplan-Meier curves, Log-rank test, χ^2 test	OSP: (High-risk vs. low-risk) $HR = 1.940$, $CI = 1.190–3.163$, $p = 0.008$
Mendiola et al. [103]	Progression-Free Survival, Overall Survival	Kaplan-Meier curves, Log-rank test, AIC-based model selection, ROC curves, Cox proportional hazards regression	OS: sensitivity = 87.2%, specificity = 86.4%
Gevaert et al. [50]	Platin Resistance/Sensitivity, Stage	Principal component analysis, Least squares support vector machines	Platin-Resistance/Sensitivity: sensitivity = 67%, specificity = 40%, accuracy = 51.11%
Bachvarov et al. [12]	Chemoresistance	Hierarchical Clustering, Support vector machines	No prediction metric applied
Netinatsunthorn et al. [115]	Overall Survival, Recurrence-Free Survival	Kaplan-Meier curves, Cox proportional hazards regression	OS: $HR = 1.98$, 95% $CI = 1.28–3.79$, $p = 0.0138$; RFS: $HR = 3.36$, 95% $CI = 1.60–7.03$, $p = 0.0017$
De Smet et al. [35]	Stage I vs. Advanced stage, Platin-sensistive vs. Platin-resistant	Principal component analysis, Least squares support vector machines	Estimated Classification Accuracy: Stage I vs Advanced Stage = 100%, Platin-sensitive vs. Platin-resistant = 76.9%

Table A.6: (continued)

Helleman et al. [66]	Chemoresponse (responder vs. non-responder)	Class prediction, Hierarchical clustering, Principal component analysis	Test set: $PPV = 24\%$, $NPV = 97\%$, sensitivity = 89%, specificity = 59%
Spentzos et al. [148]	Chemoresponse (pathological-CR or PD), Disease-Free survival, Overall Survival	Class prediction analysis, Compound covariate algorithm, Average linkage hierarchical clustering, Kaplan-Meier curves, Log-rank test, Cox proportional hazards regression	Cox PH (resistant vs. sensitive): Recurrence $HR = 2.7$ (95% $CI = 1.2$–6.1), Death $HR = 3.9$ (95% $CI = 3.1$–11.4)
Jazaeri et al. [74]	Clinical response	Class prediction	9 most significantly differentially expressed genes, primary chemoresistant vs. primary chemosensitive: accuracy = 77.8%
Raspollini et al. [128]	Overall Survival (high vs. low)	Univariate logistic regression, χ^2 test	COX-2: $OR = 0.23$, 95% $CI = 0.06$–0.77, $p = 0.017$; MDR1: $OR = 0.01$, 95% $CI = 0.002$–0.09, $p =< 0.0005$

Table A.6: (continued)

Hartmann et al. [63]	Time To Relapse (early vs.late)	Support vector machine, Kaplan-Meier curves, Log-rank test, average linkage clustering	Accuracy = 86%, PPV = 95%, $NPV = 67\%$
Spentzos et al. [147]	Disease-Free Survival, Overall Survival	Supervised pattern recognition/class prediction, Kaplan-Meier curves, Log-rank test, Cox proportional hazards regression	Unfavourable vs. Favourable OS : (CPH) $HR = 4.6$, 95% $CI = 2.0$–10.7, $p = 0.0001$
Selvanayagam et al. [142]	Chemoresistance (chemoresistant vs. chemosensitive)	Supervised voice-pattern recognition algorithm (clustering)	$PPV = 1$, $NPV = 1$
Iba et al. [72]	Chemoresponse, Overall Survival	Kaplan-Meier curves, Log-rank test, Cox propotionate hazards regression, ROC analysis, χ^2 test, Student's T test, Mann-Whitney U test	'Patients with c-myc expression of over 200 showed a significantly better 5-year survival rate (69.8% vs. 43.5%)', $p < 0.05$
Kamazawa et al. [77]	Chemoresponse (CR or PR vs. NC or PD)	Defined threshold expression to divide responders and non-responders	MDR-1 (all samples): specificity = 95%, sensitivity = 100%, predictive value = 96%

Table A.6: (continued)

Vogt et al. [161]	Chemoresistance	Correlation of AUC from in-vitro ATP-CVA and gene expression	All p values for correlation of drugs and genes were > 0.05

Table A.7: List of genes reported by studies included in this review. Gene names have been standardised. Genes in bold were selected by more than two studies.

A1BG	CST6	HRASLS	NID1	SHFM1
A2M	CST9L	Hs.120332	NIT1	SHOX
AADAC	CT45A6	HS3ST1	NKIRAS2	SIDT1
AAK1	CTA-246H3.1	HS3ST5	NKX31	SIGLEC8
ABCA13	CTNNBL1	HSD11B2	NKX62	SIRT5
ABCA4	CTSD	HSD17B11	NLGN1	SIRT6
ABCB1	CUTA	HSPA1L	NOP5/58	SIVA1
ABCB10	CX3CL1	HSPA4	NOS3	SIX2
ABCB11	CXCL1	HSPA8	NOTCH4	SKA3
ABCB7	CXCL10	HSPB7	NOV	SLAMF7
ABCC3	CXCL12	HSPD1	NOX1	SLC12A2
ABCC5	CXCL13	HTATIP2	NPAS3	SLC12A4
ABCD2	CXCR4	HTN1	NPR1	SLC14A1
ABCG2	CYB5B	HTR3A	NPR3	SLC15A2
ABLIM1	CYBRD1	ICAM1	NPTX2	SLC1A1
ACADVL	CYP27A1	ICAM5	NPTXR	SLC1A3
ACAT2	CYP2E1	ID1	NPY	SLC22A5
ACKR2	CYP3A7	ID4	NRBP2	SLC25A37
ACKR3	CYP4X1	IDI1	NRG4	SLC25A41
ACO2	CYP4Z1	IFIT1	NRP1	SLC25A5
ACOT13	CYP51A1	IGF1R	NSFL1C	SLC26A9
ACP1	CYSTM1	IGFBP2	NSL1	SLC27A6
ACRV1	CYTH3	IGFBP5	NSMCE4A	SLC29A1
ACSM1	D4S234E	IGHM	NT5C3A	SLC2A1
ACSS3	DAP	IGKC	NTAN1	SLC2A5
ACTA2	DAPL1	IGKV1-5	NTF4	SLC37A4
ACTB	DBI	IHH	NUDT21	SLC39A2
ACTBL3	DCBLD2	IKZF4	NUDT9	SLC4A11
ACTG2	DCHS1	IL11RA	NUS1	SLC5A1
ACTR3B	DCK	IL15	OAS3	SLC5A3
ACTR6	DCTN5	IL17RB	OASL	SLC5A5
ADAMDEC1	DCTPP1	IL1B	ODF4	SLC6A3
ADAMTS5	DCUN1D4	IL23A	OGFOD3	SLC7A2
ADIPOR2	DCUN1D5	IL27	OGN	SMAD2

Table A.7: (continued)

ADK	DDB1	IL6	OPA3	SMC4
AEBP1	DDB2	IL8	OR10A3	SMG1
AF050199	DDR1	IMPA2	OR2AG1	SMPD2
AF052172	DDX23	ING3	OR4C15	SNIP1
AFM	DDX49	INHBA	OR51B5	SNRPA1
AFTPH	DEFB132	INPP5A	OR51I1	SNRPC
AGFG1	DERL1	INPP5B	OR6F1	SNRPD3
AGR2	DFNB31	INSR	OR9G9	SNX13
AGT	DHCR7	INTS12	OSGEPL1	SNX19
AIPL1	DHRS11	INTS9	OSGIN2	SNX7
AKAP12	DHRS9	IRF2BP1	OSM	SOAT2
AKR1A1	DHX15	ISCA1	OXTR	SOBP
AKR1C1	DHX29	ISG20	P2RX4	SORBS3
AKT1	DIAPH3	ITGAE	PABPC4	SOS1
AKT2	DICER1	ITGB2	PAGR1	SOX12
ALCAM	DIRC1	ITGB6	PAH	SOX21
ALDH5A1	DKK1	ITGB7	PAK4	SPANXD
ALDH9A1	DLAT	ITLN1	PALB2	SPATA13
ALG5	DLEU2	ITM2A	PARD6B	SPATA18
ALMS1	DLG1	ITM2C	PAX6	SPATA4
AMPD1	DLG3	ITPR2	PBK	SPC25
ANKHD1	DLGAP4	ITPRIP	PBX2	SPDEF
ANKRD27	DLGAP5	JAG2	PBXIP1	SPEN
ANXA3	DMRT3	JAK2	PCF11	SPHK2
ANXA4	DNAH2	JAKMIP2	PCGF3	SPOCK2
AOC1	DNAH7	KCNB1	PCK1	SPTBN2
AP2A2	DNAJB12	KCNE3	PCNA	SRC
APC	DNAJB5	KCNH2	PCNXL2	SREBF2
API5	DNAJC16	KCNJ16	PCOLCE	SRF
APOE	DNASE1L3	KCNN1	PCSK6	SRRM1
AQP10	DOCK3	KCNN3	PDCD2	SRSF3
AQP5	DPH2	KCTD1	PDE3A	SSR1
AQP6	DPM1	KCTD5	PDGFA	SSR2
AQP9	DPP7	KDELC1	PDGFRA	SSUH2
ARAF	DPYSL2	KDELR1	PDGFRB	SSX2IP

Table A.7: (continued)

ARAP1	DRD4	KDELR2	PDP1	ST6GALNAC1
AREG	DTYMK	KDM4A	PDSS1	STC2
ARFGEF2	DUSP2	Ki67	PDZK1	STK38
ARHGAP29	DUSP4	KIAA0125	PEBP1	STX12
ARHGDIA	DUX3	KIAA0141	PEX11A	STX1B
ARL14	DYNLT1	KIAA0226	PEX6	STX7
ARL6IP4	DYRK3	KIAA0368	PFAS	STXBP2
ARMC1	E2F2	KIAA1009	PGAM1	STXBP6
ARNT2	ECH1	KIAA1033	PHF3	SUB1
ARPC4	EDF1	KIAA1324	PHGDH	SULT1C2
ASAP1	EDN1	KIAA1551	PHKA1	SULT2B1
ASAP3	EDNRA	KIAA2022	PHKA2	SUPT5H
ASF1A	EDNRB	KIAA4146	PI3	SUSD4
ASIP	EEF1A2	KIF3A	PIC3CD	SUV420H1
ASPA	EFCAB14	KIFC3	PIGC	SV2C
ASPHD1	EFEMP2	KIT	PIGR	SYNM
ASS1	EFNB2	KLF12	PIK3CG	SYT1
ASUN	EGF	KLF5	PIP5K1B	SYT11
ATM	EGFR	KLHDC3	PITRM1	SYT13
ATP1B3	EHD1	KLHL7	PKD1	TAC3
ATP5D	EHF	KLK10	PKHD1	TAP1
ATP5F1	EI24	KLK6	PLA2G7	TASP1
ATP5L	EIF1	KPNA3	PLAA	TBCC
ATP6V0E1	EIF2AK2	KPNA6	PLAU	TBP
ATP7B	EIF3K	KRT10	PLAUR	TCF15
ATP8A2	EIF4E2	KRT12	PLCB3	TCF7L2
AUP1	EIF5	KYNU	PLEC	TENM3
AURKA	ELF3	L1TD1	PLEK	TEX30
AURKC	ELF5	LAMB1	PLIN2	TFF1
AVIL	EML4	LAMTOR5	PLS1	TFF3
B3GALNT1	ENC1	LARP4	PMM1	TFPI2
B3GNT2	ENOPH1	LAX1	PMP22	TGFB1
B4GALT5	ENSA	LAYN	PMVK	THBS4
BAG3	ENTPD4	LBR	PNLDC1	TIAM1
BAIAP2L1	EPB41L4A	LCMT2	PNLIPRP2	TIMM10B

Table A.7: (continued)

BAK1	EPCAM	LCTL	PNMA5	TIMM17B
BASP1	EPHB2	LDB1	POFUT2	TIMP1
BAX	EPHB3	LDHB	POLH	TIMP2
BCHE	EPHB4	LGALS4	POLR3K	TIMP3
BCL2A1	EPOR	LGR5	POMP	TKTL1
BCL2L11	ERBB3	LHB	POU2AF1	TLE2
BCL2L12	ERCC8	LHX1	POU5F1	TM9SF2
BCR-ABL	ERMP1	LIN28A	PPAP2B	TM9SF3
BEAN	ESF1	LINGO1	PPAT	TMCC1
BEST4	ESM1	LIPA	PPCDC	TMED5
BFSP1	ESR1	LIPC	PPCS	TMEM139
BFSP2	ESRP2	LIPG	PPFIA3	TMEM14B
BGN	ESYT1	LMO3	PPIC	TMEM150A
BHLHE40	ETS1	LMO4	PPIE	TMEM161A
BIN1	ETV1	LOC100129250	PPP1R1A	TMEM259
BIRC5	EVA1A	LOC149018	PPP1R1B	TMEM260
BIRC6	EXOC6B	LOC1720	PPP1R2	TMEM45A
BLCAP	EXTL1	LOC389677	PPP1R26	TMEM50A
BLMH	EYA2	LOC642236	PPP2R3C	TMPRSS3
BMP8B	F2R	LOC646808	PPP2R5C	TMSB15B
BMPR1A	FAAH	LOC90925	PPP2R5D	TMTC1
BNIP3	FABP1	LPAR6	PPP4R4	TMX2
BOLA3	FABP7	LPCAT2	PPP6R1	TNFRSF17
BPTF	FADS1	LPCAT4	PRAP1	TNS1
BRCA1	FADS2	LPHN2	PRELP	TOMM40
BRCA2	FAM133A	LRIG1	PRKAB1	TONSL
BRSK1	FAM135A	LRIT1	PRKCH	TOP1
BTN3A3	FAM155B	LRRC16B	PRKCI	**TOP2A**
BTNL9	FAM174B	LRRC17	PRKD3	TOX3
C11orf16	FAM19A4	LRRC59	PROC	**TP53**
C11orf74	FAM211B	LRSAM1	PROK1	TP53TG5
C12orf5	FAM217B	LSAMP	PRPF31	TP73
C16orf89	FAM49B	LSM14A	PRRX1	TPD52
C17orf45	FAM8A1	LSM3	PRSS16	TPM2
C17orf53	FANCB	LSM7	PRSS22	TPP2

Table A.7: (continued)

C17orf70	FANCE	LSM8	PRSS3	TPPP
C1orf109	FANCF	LTA4H	PRSS36	TPRKB
C1orf115	FANCG	LTB	PSAT1	TRA
C1orf159	FANCI	LTK	PSMB5	TRAF3IP2
C1orf198	FARP1	LUC7L2	PSMB9	TRAM1
C1orf27	FAS	LY6K	PSMC4	TRAPPC4
C1orf68	FASLG	LY96	PSMD1	TRAPPC9
C1QTNF3	FBXL18	LZTFL1	PSMD12	TREML1
C20orf199	FCGBP	MAB21L2	PSMD14	TREML2
C2orf72	FCGR3B	MAD2L2	PSME4	TRIAP1
C4A	FEN1	MAGEE2	PTBP1	TRIM27
C4BPA	FEZ1	MAGEF1	PTCH2	TRIM49
C6orf120	FGF2	MAK	PTEN	TRIM58
C6orf124	FGFBP1	MAMLD1	PTGDS	TRIML2
C9orf3	FGFR1OP	MANF	PTGS2	TRIT1
C9orf47	FGFR1OP2	MAP6D1	PTP4A1	TRMT1L
CA13	FGFR2	MAPK1	PTP4A2	TRO
CACNA1B	FHL2	MAPK1IP1L	PTPRN2	TRPV4
CACNG6	FILIP1	MAPK3	PTPRS	TRPV6
CADM1	FJX1	MAPK8IP3	PWP2	TSPAN3
CALML3	FKBP11	MAPK9	QPRT	TSPAN4
CAMK2B	FKBP1B	MAPKAP1	R3HDM2	TSPAN6
CAMK2N1	FKBP7	MAPKAPK2	RAB26	TSPAN7
CANX	FLII	MARCKS	RAB27B	TSR1
CAP1	FLJ41501	MARK4	RAB40B	TTC31
CAP2	FLNC	MATK	RAB5B	TTLL6
CAPN13	FLOT2	MB	RAB5C	TTPAL
CAPN5	FLT1	MBOAT7	RABIF	TTYH1
CASC3	FMN2	MCF2L	RAC1	TUBB3
CASP9	FMO1	MCL1	RAC3	TUBB4A
CASS4	FN1	MCM3	RAD23A	TUBB4Q
CATSPERD	**FOXA2**	MDC1	RAD51	TUSC3
CC2D1A	FOXD4L2	MDFI	RAD51AP1	UBD
CCBL1	FOXJ1	MDK	RANBP1	UBE2I
CCDC130	FOXO3	MDR-1	RANGAP1	UBE2K

Table A.7: (continued)

CCDC135	FSCN1	MEA1	RARRES2	UBE2L3
CCDC147	FXYD6	MEAF6	RB1	UBE4B
CCDC167	FZD4	MECOM	RBBP7	UBR5
CCDC19	FZD5	MEF2B	RBFA	UGT2B17
CCDC53	G0S2	MEGF11	RBM11	UGT8
CCDC9	G3BP1	MEST	RBM39	UHRF1BP1
CCL13	GABRP	METRN	RCHY1	UMOD
CCL2	GAD1	METTL13	RER1	UPK1A
CCL28	GALNT10	METTL4	RFC3	UPK1B
CCM2L	GAP43	MFAP2	RGL2	UQCRC2
CCNA2	GART	MFSD7	RGP1	URI1
CCNG2	GATAD2A	MGMT	RGS19	USP14
CCT6A	GCH1	MINOS1	RHOT1	USP18
CCZ1	GCHFR	MKRN1	RHPN2	USP21
CD34	GCM1	MLF2	RIIAD1	UST
CD38	GDF6	MLH1	RIN1	UTP11L
CD44	GFRA1	MLX	RIT1	UTP20
CD46	GGCT	MMP1	RNF10	UVRAG
CD70	GGT1	MMP10	RNF13	VDR
CD97	GJB1	MMP12	RNF14	VEGFA
CDC42EP4	GLRX	MMP13	RNF148	VEGFB
CDCA2	GMFB	MMP16	RNF34	VEZF1
CDH12	GMPR	MMP17	RNF6	VPS39
CDH19	GNA11	MMP3	RNF7	VPS52
CDH3	GNAO1	MMP7	RNF8	VPS72
CDH4	GNAZ	MMP9	RNGTT	VTCN1
CDH5	GNG4	MPZL1	RNPEPL1	VTI1B
CDK17	GNG7	MRPL2	ROBO1	WBP2
CDK20	GNL2	MRPL35	ROR1	WBP4
CDK5R1	GNMT	MRPL49	ROR2	WDR12
CDK8	GNPDA1	MRPS12	RP13-347D8.3	WDR45B
CDKN1A	GOLPH3	MRPS17	RP13-36C9.6	WDR7
CDY1	GPIHBP1	MRPS24	RPA3	WDR77
CDYL2	GPM6B	MRPS9	RPL23	WIT1
CEACAM5	GPR137	MRS2	RPL29P17	WIZ

Table A.7: (continued)

CEACAM6	GPT2	MSH2	RPL31	WNK4
CEACAM7	GPX2	MSL1	RPL36	WNT16
CEP55	GPX3	MSMO1	RPP30	WT1
CES1	GPX8	MST1	RPS15	WTAP
CES2	GRAMD1B	MT1G	RPS16	WWOX
CFI	GRB2	MTCP1	RPS19BP1	XBP1
CH25H	GRK6	MTMR11	RPS24	XPA
CHIT1	GRM2	MTMR2	RPS28	XPO4
CHPF2	GRPEL1	MTPAP	RPS4Y1	XYLT1
CHRDL1	GRSF1	MTUS1	RPS6KA2	Y09846
CHRNE	GSPT1	MTX1	RPSA	YBX1
CHST6	GSTM2	MUS81	RRAGC	YIPF3
CHTOP	GSTT1	**MUTYH**	RRBP1	YIPF6
CIAPIN1	GTF2E1	MXD1	RRN3	YLPM1
CIB1	GTF2F2	MXI1	RSL24D1	YWHAE
CIB2	GTF2H5	MYBPC1	RSU1	YWHAZ
CIITA	GTPBP4	MYC	RTN4R	ZBTB11
CILP	GUCY1B3	MYCBP	RXRB	ZBTB16
CITED2	GYG1	MYL9	RYBP	ZBTB8A
CKLF	GYPC	MYO1D	RYR3	ZC3H13
CLCA1	GZMB	MYOM1	S100A10	ZCCHC8
CLCNKB	GZMK	NANOS1	S100A4	ZEB2
CLDN10	H2AFX	NASP	S100P	ZFHX4
CLIP1	H3F3A	NBEA	SAMD4B	ZFP91
CNDP1	HAP1	NBL1	SASH1	ZFR2
CNKSR3	HBG2	NBN	SCAMP3	ZKSCAN7
CNN2	HDAC1	NCAM1	SCARF1	ZMYND11
CNOT8	HDAC2	NCAPD2	SCG2	ZNF106
CNTFR	HECTD4	NCAPG	SCGB1C1	ZNF12
cofilin1	HES1	NCAPH	SCGB3A1	ZNF124
COL10A1	HEY1	NCKAP5	SCNM1	ZNF148
COL21A1	HHIPL2	NCOA1	SCO2	ZNF155
COL3A1	HIF1A	NCOR2	SCUBE2	ZNF180
COL4A4	HIP1R	NCR2	SDF2L1	ZNF200
COL4A6	HIPK1	NCSTN	SEC14L2	ZNF292

Table A.7: (continued)

COL6A1	HIST1H1C	NDRG2	SELT	ZNF337
COL7A1	HK2	NDST1	SEMA3A	ZNF432
COX8A	HLAA	NDUFA12	SENP3	ZNF467
CPD	HLADMB	NDUFA9	SENP6	ZNF48
CPE	HLADOB	NDUFAB1	SEPN1	ZNF503
CPEB1	HMBOX1	NDUFAF4	SERPINB6	ZNF521
CRCT1	HMGCS1	NDUFB4	SERPIND1	ZNF569
CREB5	HMGCS2	NDUFS5	SERPINF1	ZNF644
CRYAB	HMGN1	NEBL	SERTAD4	ZNF71
CRYBB1	HMOX2	NETO2	SETBP1	ZNF711
CRYL1	HNRNPA1	NEUROD2	SF3A3	ZNF74
CRYM	HNRNPUL2	NFE2	SF3B4	ZNF76
CSE1L	HOPX	NFE2L3	SGCB	ZNF780B
CSPP1	HOXA5	NFIB	SGCG	ZYG11A
CSRP1	HOXB6	NFKBIB	SGPP1	
CSRP3	HPN	NFS1	SH3PXD2A	

Table A.8: Genes chosen most commonly by studies in review, listed by number of papers selecting each gene. Gene function and links to cancer obtained via cursory literature search.

Gene Symbol	Number of studies	Function	Expression links to cancer in literature
AGR2	4	Cell migration and growth	Prostate, breast, ovarian, pancreatic
MUTYH	3	Oxidative DNA damage repair	Colorectal
AKAP12	3	Subcellular compartmentation of PKA	Colorectal, lung, prostate
TP53	3	Cell cycle regulation	Breast
TOP2A	3	Required for DNA replication	Breast, prostate, ovarian
FOXA2	3	Liver-specific transcription factor	Lung, prostate

Table A.8: (continued)

SRC	2	Regulation of cell growth	Colon, liver, lung, breast, pancreatic
SIVA1	2	Pro-apoptotic protein	Many cancers
ALDH9A1	2	Aldehyde dehydrogenase	Many cancers
LGR5	2	Associated with stem cells	Cancer stem cells
EHF	2	Epithelial differentiation and proliferation	Prostate
BAX	2	Apoptotic activator	Colon, breast, prostate, gastric, leukaemia
CES2	2	Intestine drug clearance	Colorectal
CPE	2	Synthesis of hormones and neurotransmitters	
FGFBP1	2	Cell proliferation, differentiation and migration	Colorectal, pancreatic
TUBB4A	2	Component of microtubules	
ZNF12	2	Transcription regulation	
RBM39	2	Steroid hormone receptor-mediated transcription	
RFC3	2	Required for DNA replication	
GNPDA1	2	Triggers calcium oscillations in mammalian eggs	
ANXA3	2	Regulation of cellular growth	Prostate, ovarian
NFIB	2	Activates transcription and replication	Breast
ACTR3B	2	Actin cyctoskeleton organisation	Lung
YWHAE	2	Mediates signal transduction	Lung, endometrial

Table A.8: (continued)

CYP51A1	2	Drug metabolism and lipid synthesis	
HMGCS1	2	Cholesterol synthesis and ketogenesis	
ZMYND11	2	Transcriptional repressor	
FADS2	2	Regulates unsaturation of fatty acids	
SNX7	2	Family involved in intracellular trafficking	
ARHGDIA	2	Regulates the GDP/GTP exchange reaction of the Rho proteins	Prostate, lung,
NDST1	2	Inflammatory response	Prostate, breast
AOC1	2	Catalyses degredation of such as histamine and spermidine	
DAP	2	Positive mediator of programmed cell death	
ERCC8	2	Transcription-coupled nucleotide excision repair	
GUCY1B3	2	Catalyzes conversion of GTP to the second messenger cGMP	
HDAC1	2	Control of cell proliferation and differentiation	Prostate, breast, colorectal, gastric
HDAC2	2	Transcriptional regulation and cell cycle progression	Cervical, gastric, colorectal
IGFBP5	2	Cell proliferation, differentiation, survival, and motility	Breast

Table A.8: (continued)

IL6	2	Transcriptional inflammatory response, B cell maturation	Many cancers
LSAMP	2	Neuronal surface glycoprotein	Osteosarcoma
MDK	2	Cell growth, migration, angiogenesis	Many cancers
MYCBP	2	Stimulates the activation of E box-dependent transcription	
S100A10	2	Transport of neurotransmitters	Colorectal, lung, breast
SLC1A3	2	Glutamate transporter	
NCOA1	2	Stimulates hormone-dependent transcription	Breast, prostate
TIAM1	2	Modulates the activity of Rho GTP-binding proteins	Many cancers
VEGFA	2	Angiogenesis, cell growth, cell migration, apoptosis	Many cancers
RPL36	2	Component of ribosomal 60S subunit	
LBR	2	Anchors lamina and heterochromatin to the nuclear membrane	
ABCB1	2	ATP-dependent drug efflux pump for xenobiotic compounds	Many cancers
FASLG	2	Required for triggering apoptosis in some cell types	Many cancers
TIMP1	2	Extracellular matrix, proliferation, apoptosis	Many cancers

Table A.8: (continued)

FN1	2	Cell adhesion, motility, migration processes	Many cancers
TGFB1	2	Proliferation, differentiation, adhesion, migration	Prostate, breast, colon, lung, bladder
XPA	2	DNA excision repair	Many cancers
ABCB10	2	Mitochondrial ATP-binding cassette transporter	
POLH	2	Polymerase capable of replicating UV-damaged DNA for repair	
ITGAE	2	Adhesion, intestinal intraepithelial lymphocyte activation	
ZNF200	2	Zinc finger protein	
COL3A1	2	Collagen type III, occurring in most soft connective tissues	
ACKR3	2	G-protein coupled receptor	
EPHB3	2	Mediates developmental processes	Lung, colorectal
NBN	2	Double-strand DNA repair, cell cycle control	
PCF11	2	May be involved in Pol II release following polymerisation	
DFNB31	2	Sterocilia elongation, actin cystoskeletal assembly	
BRCA2	2	Double-strand DNA repair	Breast, ovarian
AADAC	2	Arylacetamide deacetylase	

Table A.8: (continued)

CD38	2	Glucose-induced insulin secretion	Leukaemia
CHIT1	2	Involved in degradation of chitin-containing pathogens	
CXCR4	2	Receptor specific for stromal-derived-factor-1	Breast, glioma, kidney, prostate
EFNB2	2	Mediates developmental processes	
MECOM	2	Apoptosis, development, cell differentiation, proliferation	Leukaemia
FILIP1	2	Controls neocortical cell migration	Ovarian
HSPB7	2	Heat shock protein	
LRIG1	2	Regulator of signaling by receptor tyrosine kinases	Glioma
MMP1	2	Breakdown of extracellular matrix	Gastric, breast
PSAT1	2	Phosphoserine aminotransferase	
SDF2L1	2	Part of endoplasmic reticulum chaperone complex	
TCF15	2	Regulation of patterning of the mesoderm	
EPHB2	2	Contact-dependent bidirectional signaling between cells	Colorectal
ETS1	2	Involved in stem cell development, cell senescence and death	Many cancers

Table A.8: (continued)

TRIM27	2	Male germ cell differentiation	Ovarian, endometrial, prostate
MARK4	2	Mitosis, cell cycle control	Glioma
B4GALT5	2	Biosynthesis of glycoconjugates and saccharides	

Appendix B

Chapter 3 Supplementary Information

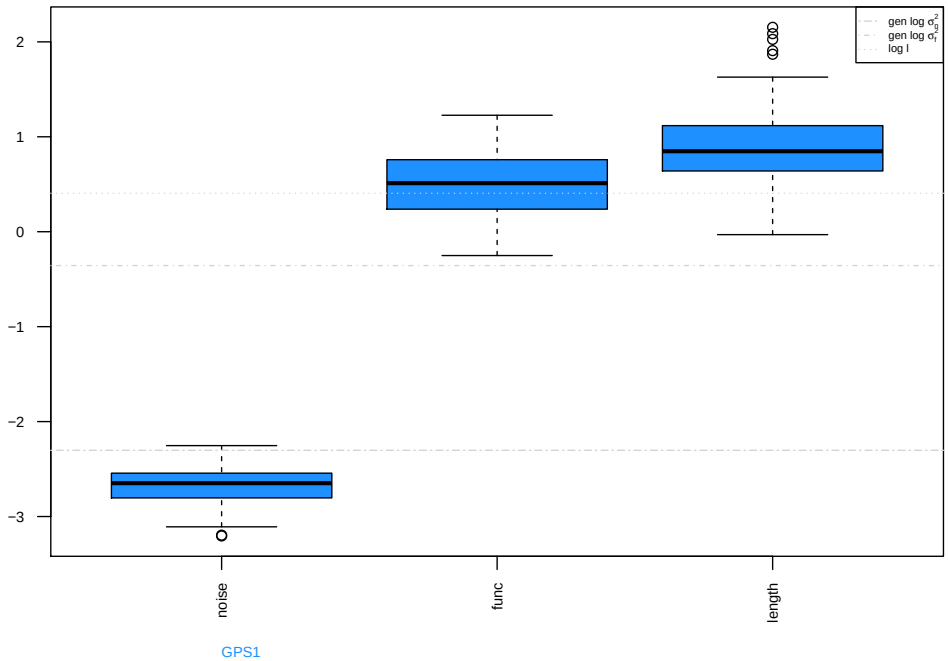


Figure B.1: Boxplots of hyperparameter values selected by GPS1 fitted using 100 generated synthetic data sets. The generating hyperparameters are marked in grey.

Appendix C

Chapter 4 Supplementary Information

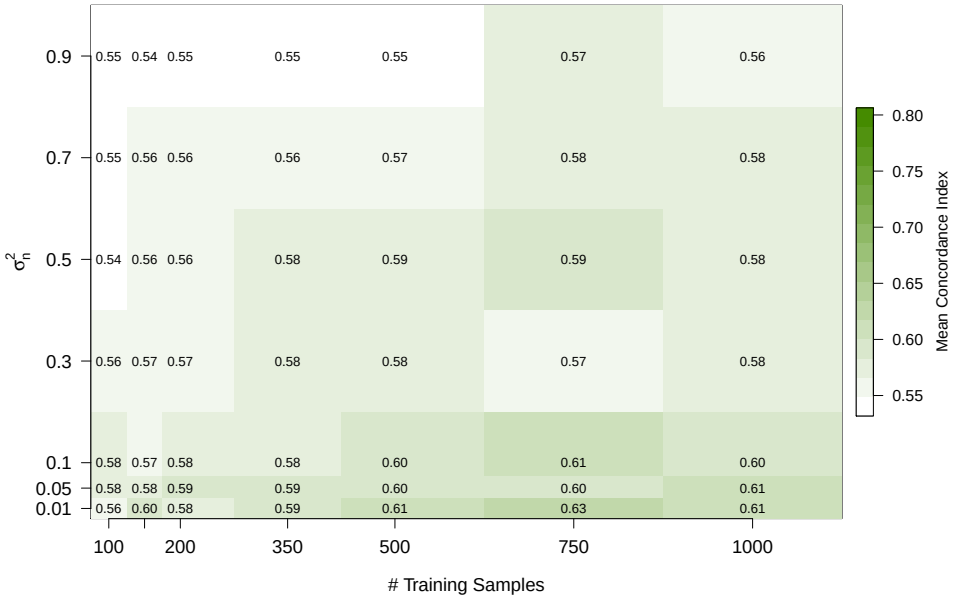
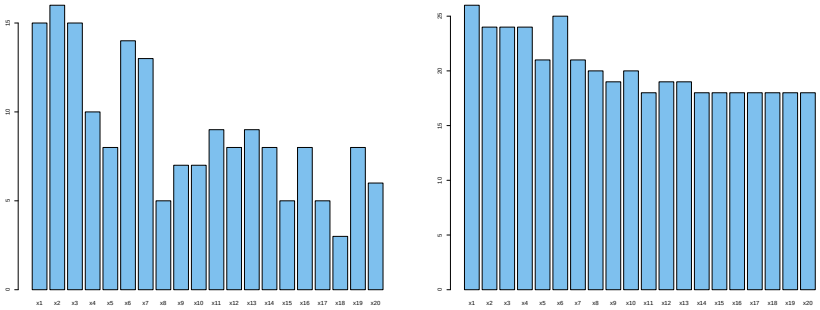


Figure C.1: Experiment 3. Mean concordance index values as generating noise variance hyperparameter and number of training samples are varied. Fitted using Coxph

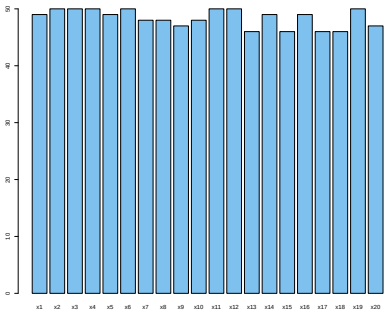
Appendix D

Chapter 5 Supplementary Information

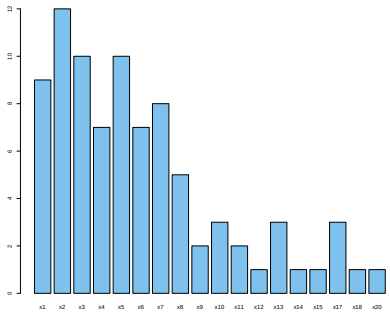


(a) FilterCoxph1

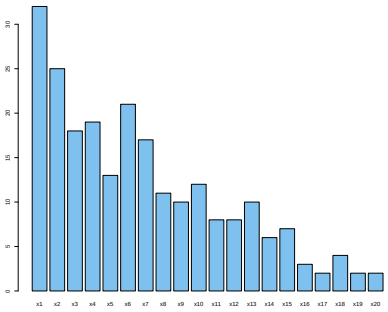
(b) FilterCoxph2



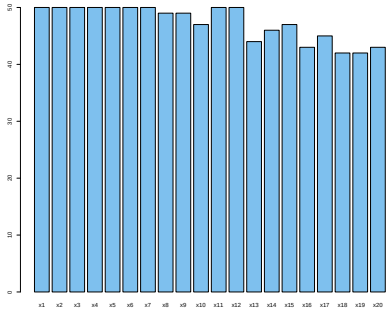
(c) Coxnet



(d) StepCoxph

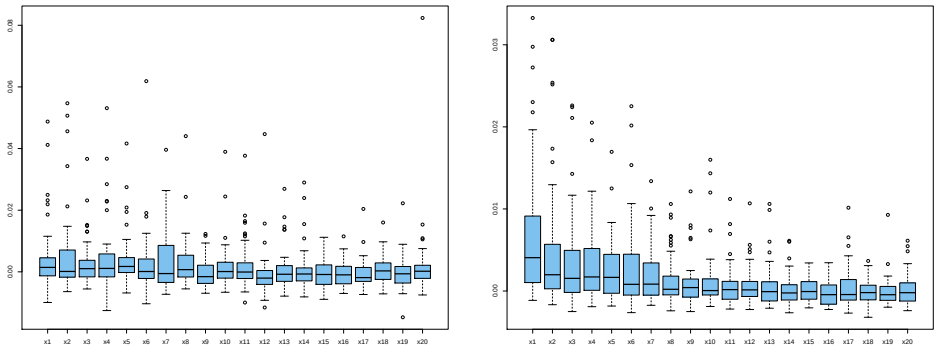


(e) GPS3BICForward



(f) GPS3BICBackward

Figure D.1: Experiment 1IR. Frequency plots of variables selected during feature selection for a range of models. a): StepCoxph b): FilterCoxph1 c): FilterCoxph2 d): Glmnet e): GPSBICBackwards f): GPSBICForwards



(a) RF (b) RSF

Figure D.2: Experiment 1IR. Boxplots of variable importance reported for each feature of random forest models. a): RF b): RSF

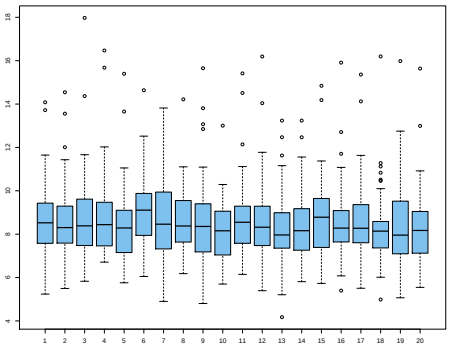
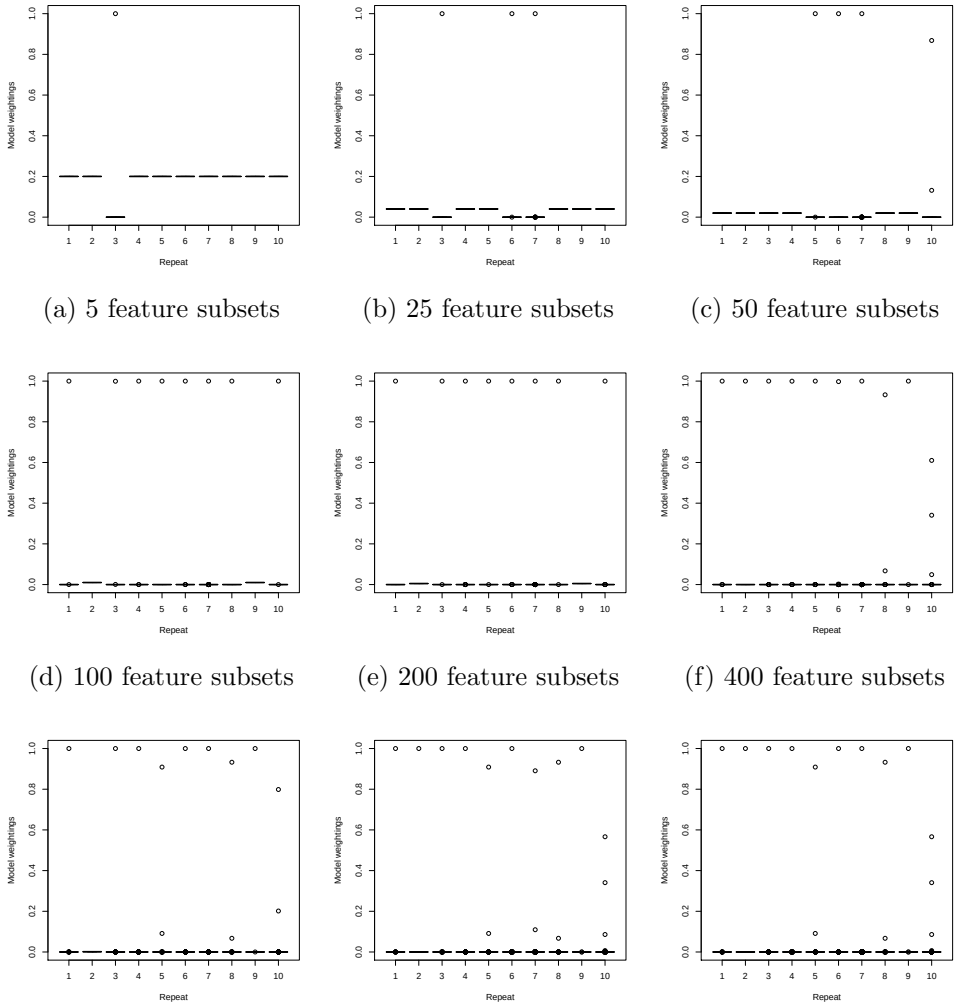


Figure D.3: Experiment 1IR. Boxplots of marginalised probability reported for each feature of GPS3SqExpRSFS.



(a) 5 feature subsets (b) 25 feature subsets (c) 50 feature subsets

(d) 100 feature subsets (e) 200 feature subsets (f) 400 feature subsets

(g) 600 feature subsets (h) 800 feature subsets (i) 1000 feature subsets

Figure D.4: Experiment 2R. Boxplots plots of GPS3SqExpRSFS model weightings for the first 10 repeats, using varing numbers of feature subsets.

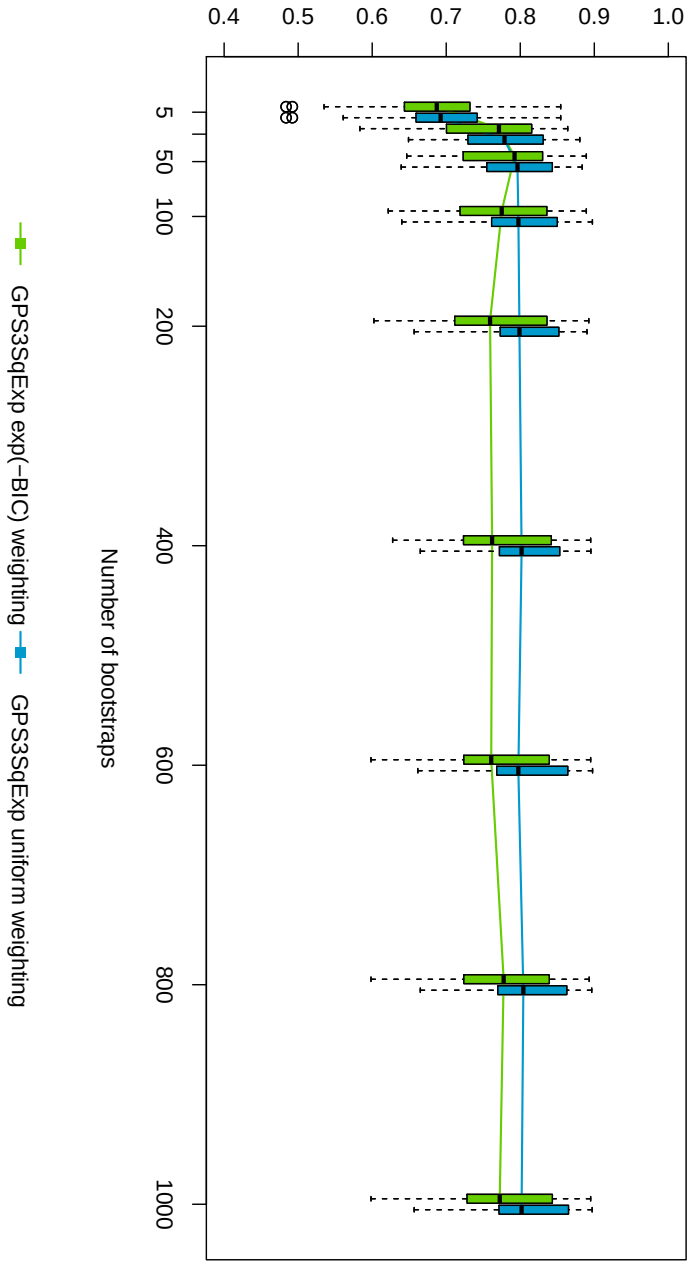


Figure D.5: Experiment 2R. Boxplots of concordance index values for ensemble predictions as the number of feature subsets are varied. Models are GPS3SqExpRSFS using exp(-BIC) wighting and GPS3SqExpRSFS using uniform weighting of models. Each boxplot represents 99 repeats. GPS3SqExpRSFS was applied to each repeat for varying numbers of feature subsets, using the specified weighting for generating ensemble predictions.

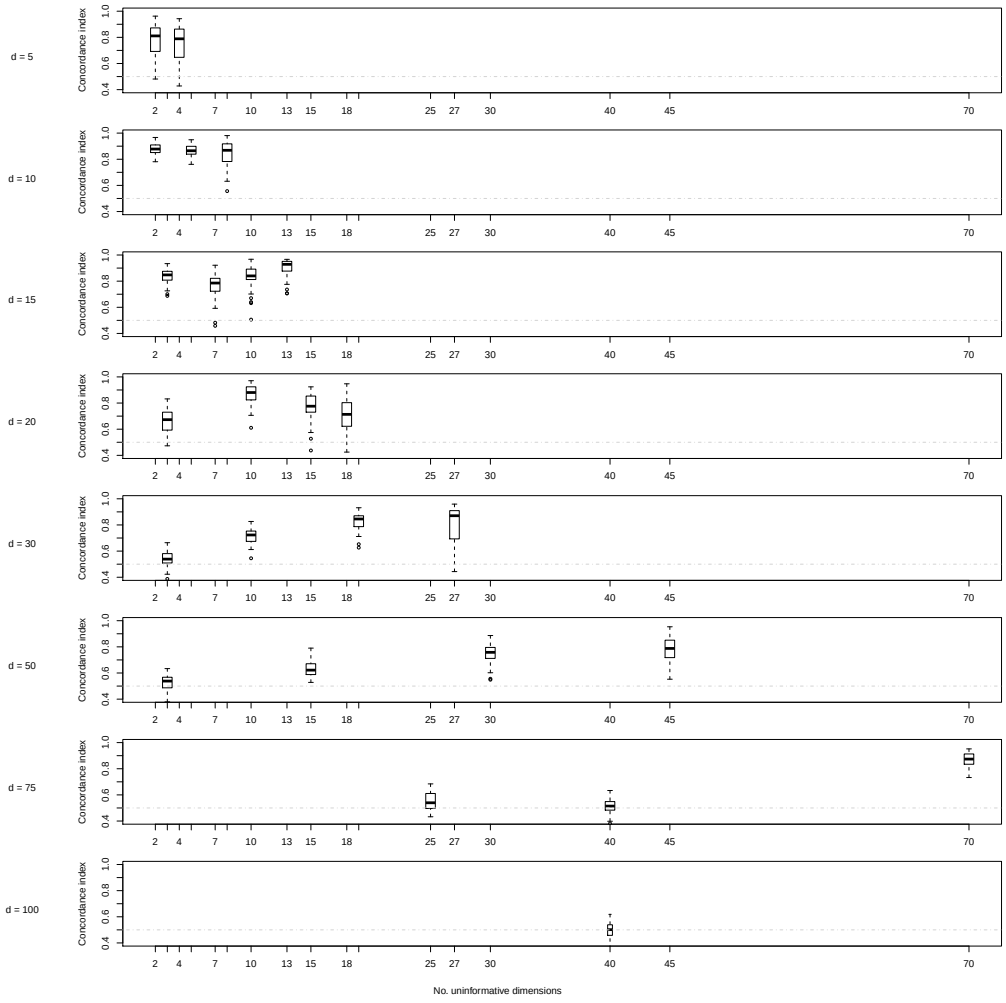


Figure D.6: Experiment 3R. Results of running GPS3 with RSFS on synthetic data with changing total number of dimensions and number non-informative dimensions. Boxplots of concordance index of the model predictions as the total dimensionality (y axis) and the number of non-informative dimensions (x axis) changes.

Bibliography

[1] Improving outcomes through personalised medicine. URL https://www.england.nhs.uk/wp-content/uploads/2016/09/improving-outcomes-personalised-medicine.pdf.

[2] Cancer Research UK: Ten-year survival over time. URL http://www.cancerresearchuk.org/prod_consump/groups/cr_common/@nre/@sta/documents/generalcontent/surv_10yrtrends_selcancers_xls.xls.

[3] The Oxford 2011 levels of evidence (OCEBM Levels of Evidence Working Group). URL http://www.cebm.net/index.aspx?o=5653.

[4] National Cancer Institute: Tumor grade, . URL https://www.cancer.gov/about-cancer/diagnosis-staging/prognosis/tumor-grade-fact-sheet.

[5] National Cancer Institute dictionary of cancer terms: Prognosis, . URL https://www.cancer.gov/publications/dictionaries/cancer-terms?cdrid=45849.

[6] National Cancer Institute: Staging, . URL https://www.cancer.gov/about-cancer/diagnosis-staging/staging.

[7] Oncotype DX®: Underlying technology, . URL http://breast-cancer.oncotypedx.com/en-US/Professional-Invasive/WhatIsTheOncotypeDXBreastCancerTest/Underlying%20Technology.aspx.

[8] Oncotype DX®: What is the colon cancer test?, . URL http://colon-cancer.oncotypedx.com/en-US/Professional/WhatIsTheColonCancerTest.aspx.

[9] Oncotype DX®: Development, . URL http://prostate-cancer.oncotypedx.com/en-US/Professional/IntroducingGPS/Development.aspx.

[10] Office for National Statistics: Cancer survival in England: adults diagnosed between 2011 and 2015 and followed up to 2016. URL `https://www.ons.gov.uk/file?uri=/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datasets/cancersurvivalratescancersurvivalinenglandadultsdiagnosed/20112015/adultcancersurvivalcorrectionfinal.xls`.

[11] Douglas G Altman and Patrick Royston. Statistics notes: the cost of dichotomising continuous variables. *BMJ*, 332(7549):1080, 2006.

[12] D. Bachvarov, S. L'Esperance, I. Popa, M. Bachvarova, M. Plante, and B. Tetu. Gene expression patterns of chemoresistant and chemosensitive serous epithelial ovarian tumors with possible predictive value in response to initial chemotherapy. *Int. J. Oncol.*, 29(4):919–33, 2006.

[13] Karla V Ballman. Biomarker: predictive or prognostic? *J. Clin. Oncol.*, 33 (33):3968–3971, 2015.

[14] David Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, Cambridge, England, 2012.

[15] S Baulies, L Belin, P Mallon, C Senechal, JY Pierga, P Cottu, MP Sablin, X Sastre, B Asselain, R Rouzier, et al. Time-varying effect and long-term survival analysis in breast cancer patients treated with neoadjuvant chemotherapy. *Br. J. Cancer*, 113(1):30, 2015.

[16] Carine A Bellera, Gaëtan MacGrogan, Marc Debled, Christine Tunon de Lara, Véronique Brouste, and Simone Mathoulin-Pélissier. Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC Med. Res. Methodol.*, 10(1):20, 2010.

[17] Richard E Bellman. *Adaptive Control Processes: a Guided Tour*. Princeton University Press, Princeton, New Jersey, 1961.

[18] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. B Methodological*, pages 289–300, 1995.

[19] Carol Bernstein, Anil R. Prasad, Valentine Nfonsam, and Harris Bernstein. DNA damage, DNA repair and cancer. In Clark Chen, editor, *New Research Directions in DNA Repair*. InTech, London, England, 2013.

[20] Leo Breiman. Consistency for a simple model of random forests. Technical report, 2004.

[21] Kjersti Brenne, Dag André Nymoen, Thea Eline Hetland, Claes G. Trope', and Ben Davidson. Expression of the Ets transcription factor EHF in serous ovarian carcinoma effusions is a marker of poor survival. *Hum. Pathol.*, 43(4): 496–505, 04 2012.

[22] J. L. Brun, A. Cortez, B. Lesieur, S. Uzan, R. Rouzier, and E. Darai. Expression of MMP-2, -7, -9, MT1-MMP and TIMP-1 and -2 has no prognostic relevance in patients with advanced epithelial ovarian cancer. *Oncol. Rep.*, 27(4):1049–57, 2012.

[23] Fred Bunz. *Principles of Cancer Genetics*. Springer, Dordrecht, Netherlands, 2016.

[24] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny: Web Application Framework for R*, 2017. URL `https://CRAN.R-project.org/package=shiny`. R package version 1.0.3.

[25] S. Y. Chao, J. H. Chiang, A. M. Huang, and W. S. Chang. An integrative approach to identifying cancer chemoresistance-associated pathways. *BMC Med. Genomics*, 4:23, 2011.

[26] Bo Chen, Rui Castro, and Andreas Krause. Joint optimization and variable selection of high-dimensional Gaussian processes. *arXiv preprint arXiv:1206.6396*, 2012.

[27] Ami Citri and Yosef Yarden. EGF–ERBB signalling: towards the systems level. *Nat. Rev. Mol. Cell Biol.*, 7(7):505–516, 2006.

[28] Alexis B. Cortot and Pasi A. Jänne. *Resistance to Targeted Therapies As a Result of Mutation(s) in the Target*, pages 1–31. Humana Press, Totowa, NJ, 2011. ISBN 978-1-60761-478-4. doi: 10.1007/978-1-60761-478-4_1. URL `https://doi.org/10.1007/978-1-60761-478-4_1`.

[29] David R Cox. Regression models and life-tables. *J. Royal Stat. Soc. B Methodological*, 34(2):187–220, 1972.

[30] Ian A Cree. Designing personalised cancer treatments. *J. Controlled Release*, 172(2):405–409, 2013.

[31] Ian A Cree, Sharon Glaysher, and Alan L Harvey. Efficacy of anti-cancer agents in cell lines versus human primary tumour tissue. *Curr. Opin. Pharmacol.*, 10 (4):375–379, 2010.

[32] Anne PG Crijns, Rudolf SN Fehrmann, Steven de Jong, Frans Gerbens, Gert Jan Meersma, Harry G Klip, Harry Hollema, Robert MW Hofstra, Gerard J te Meerman, Elisabeth GE de Vries, et al. Survival-related profile, pathways, and transcription factors in ovarian cancer. *PLoS Med.*, 6(2):181, 2009.

[33] David B. Dahl. *xtable: Export Tables to LaTeX or HTML*, 2016. URL `https://CRAN.R-project.org/package=xtable`. R package version 1.8-2.

[34] Mark A Dawson and Tony Kouzarides. Cancer epigenetics: from mechanism to therapy. *Cell*, 150(1):12–27, 2012.

[35] F. De Smet, N. L. Pochet, K. Engelen, T. Van Gorp, P. Van Hummelen, K. Marchal, F. Amant, D. Timmerman, B. L. De Moor, and I. B. Vergote. Predicting the clinical behavior of ovarian cancer from gene expression profiles. *Int. J. Gynecol. Cancer*, 16 Suppl 1:147–51, 2006.

[36] Sven de Vos. *Technique of Microarrays: Microarray Platforms*, page 8–26. Cambridge University Press, Cambridge, England, 2006. doi: 10.1017/CBO9780511545849.003.

[37] C. Denkert, J. Budczies, S. Darb-Esfahani, B. Gyorffy, J. Sehouli, D. Konsgen, R. Zeillinger, W. Weichert, A. Noske, A. C. Buckendahl, B. M. Muller, M. Dietel, and H. Lage. A prognostic gene expression index in ovarian cancer - validation across different independent data sets. *J. Pathol.*, 218(2):273–80, 2009.

[38] AS Dhillon, S Hagan, O Rath, and W Kolch. MAP kinase signalling pathways in cancer. *Oncogene*, 26(22):3279–3290, 2007.

[39] Daniela Dunkler, Michael Schemper, and Georg Heinze. Gene selection in microarray survival studies under possibly non-proportional hazards. *Bioinformatics*, 26(6):784–790, 2010.

[40] Bradley Efron. The efficiency of Cox's likelihood function for censored data. *J. Am. Stat. Assoc.*, 72(359):557–565, 1977.

[41] Rebecca L Elliott and Gerard C Blobe. Role of transforming growth factor beta in human cancer. *J. Clin. Oncol.*, 23(9):2078–2093, 2005.

[42] J. S. Ferriss, Y. Kim, L. Duska, M. Birrer, D. A. Levine, C. Moskaluk, D. Theodorescu, and J. K. Lee. Multi-gene expression predictors of single drug responses to adjuvant chemotherapy in ovarian carcinoma: predicting platinum resistance. *PLoS One*, 7:e30550, 2012.

[43] Office for National Statistics. Cancer survival in England: Patients diagnosed between 2010 and 2014 and followed up to 2015. 2016. URL `https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancersurvivalinenglandadultsdiagnosed/2010and2014andfollowedupto2015`.

[44] Office for National Statistics. Cancer registration statistics, England: 2015. 2017. URL `https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancerregistrationstatisticsengland/2015`.

[45] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, 33(1): 1–22, 2010. URL `http://www.jstatsoft.org/v33/i01/`.

[46] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Ann. Stat.*, pages 1189–1232, 2001.

[47] Elena Galletti, Matteo Magnani, Michela L Renzulli, and Maurizio Botta. Paclitaxel and docetaxel resistance: molecular mechanisms and development of new generation taxanes. *ChemMedChem*, 2(7):920–942, 2007.

[48] Benjamin Frederick Ganzfried, Markus Riester, Benjamin Haibe-Kains, Thomas Risch, Svitlana Tyekucheva, Ina Jazic, Xin Victoria Wang, Mahnaz Ahmadifar, Michael J Birrer, Giovanni Parmigiani, et al. curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome. *Database*, 2013:bat013, 2013.

[49] Edward I George and Robert E McCulloch. Approaches for Bayesian variable selection. *Stat. Sin.*, pages 339–373, 1997.

[50] O. Gevaert, F. De Smet, T. Van Gorp, N. Pochet, K. Engelen, F. Amant, B. De Moor, D. Timmerman, and I. Vergote. Expression profiling to predict the clinical behaviour of ovarian cancer fails independent evaluation. *BMC Cancer*, 8:18, 2008.

[51] J. P. Gillet, J. Wang, A. M. Calcagno, L. J. Green, S. Varma, M. Bunkholt Elstrand, C. G. Trope, S. V. Ambudkar, B. Davidson, and M. M. Gottesman. Clinical relevance of multidrug resistance gene expression in ovarian serous carcinoma effusions. *Molecular Pharmaceutics*, 8(6):2080–8, 2011.

[52] J. P. Gillet, A. M. Calcagno, S. Varma, B. Davidson, M. Bunkholt Elstrand, R. Ganapathi, A. A. Kamat, A. K. Sood, S. V. Ambudkar, M. V. Seiden, B. R. Rueda, and M. M. Gottesman. Multidrug resistance-linked gene signature predicts overall survival of patients with primary ovarian serous carcinoma. *Clin. Cancer Res.*, 18:3197–206, 2012.

[53] Jean-Pierre Gillet, Anna Maria Calcagno, Sudhir Varma, Miguel Marino, Lisa J Green, Meena I Vora, Chirayu Patel, Josiah N Orina, Tatiana A Eliseeva, Vineet Singal, et al. Redefining the relevance of established cancer cell lines to the study of mechanisms of clinical anti-cancer drug resistance. *Proc. Natl. Acad. Sci. U.S.A.*, 108(46):18708–18713, 2011.

[54] S Glaysher, FG Gabriel, P Johnson, M Polak, LA Knight, Katharine Parker, Matthew Poole, A Narayanan, and IA Cree. Molecular basis of chemosensitivity of platinum pre-treated ovarian cancer to chemotherapy. *Br. J. Cancer*, 103 (5):656–662, 2010.

[55] Barbara A Goff. Advanced ovarian cancer: what should be the standard of care? *J. Gynecol. Oncol.*, 24(1):83–91, 2013.

[56] Michael M. Gottesman. Mechanisms of cancer drug resistance. *Annu. Rev. Med.*, 53(1):615–627, 2002. doi: 10.1146/annurev.med.53.082901.103929. URL `https://doi.org/10.1146/annurev.med.53.082901.103929`. PMID: 11818492.

[57] Kristian A Gray, Bethan Yates, Ruth L Seal, Mathew W Wright, and Elspeth A Bruford. Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res.*, page gku1071, 2014.

[58] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3(Mar):1157–1182, 2003.

[59] AG Hall and MJ Tilby. Mechanisms of action of, and modes of resistance to, alkylating agents used in the treatment of haematological malignancies. *Blood Rev.*, 6(3):163–173, 1992.

[60] Angela Hamblin, Sarah Wordsworth, Jilles M Fermont, Suzanne Page, Kulvinder Kaur, Carme Camps, Pamela Kaisaki, Avinash Gupta, Denis Talbot, Mark

Middleton, et al. Clinical applicability and cost of a 46-gene panel for genomic analysis of solid tumours: Retrospective validation and prospective audit in the UK National Health Service. *PLoS Medicine*, 14(2):e1002230, 2017.

[61] Y. Han, H. Huang, Z. Xiao, W. Zhang, Y. Cao, L. Qu, and C. Shou. Integrated analysis of gene expression profiles associated with response of platinum/paclitaxel-based treatment in epithelial ovarian cancer. *PLoS One*, 7:e52745, 2012.

[62] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *Cell*, 100 (1):57–70, 2000.

[63] L. C. Hartmann, K. H. Lu, G. P. Linette, W. A. Cliby, K. R. Kalli, D. Gershenson, R. C. Bast, J. Stec, N. Iartchouk, D. I. Smith, J. S. Ross, S. Hoersch, V. Shridhar, J. Lillie, S. H. Kaufmann, E. A. Clark, and A. I. Damokosh. Gene expression profiles predict early relapse in ovarian cancer after platinum-paclitaxel chemotherapy. *Clin. Cancer Res.*, 11:2149–55, 2005.

[64] Patrick J Heagerty, Thomas Lumley, and Margaret S Pepe. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344, 2000.

[65] V. Heinemann, J. Y. Douillard, M. Ducreux, and M. Peeters. Targeted therapy in metastatic colorectal cancer – an example of personalised medicine in action. *Cancer Treat. Rev.*, 39:592–601, 2013. URL `https://doi.org/10.1016/j.ctrv.2012.12.011`.

[66] J. Helleman, M. P. Jansen, P. N. Span, I. L. van Staveren, L. F. Massuger, M. E. Meijer-van Gelder, F. C. Sweep, P. C. Ewing, M. E. van der Burg, G. Stoter, K. Nooter, and E. M. Berns. Molecular profiling of platinum resistant ovarian cancer. *Int. J. Cancer*, 118(8):1963–71, 2006.

[67] Stacey L Hembruff and Nikki Cheng. Chemokine signaling in cancer: Implications on the tumor microenvironment and therapeutic targeting. *Cancer Ther.*, 7(A):254, 2009.

[68] Viviane Hess, Roger A'Hern, Nazar Nasiri, D Michael King, Peter R Blake, Desmond PJ Barton, John H Shepherd, T Ind, J Bridges, K Harrington, et al. Mucinous epithelial ovarian cancer: a separate entity requiring specific treatment. *J. Clin. Oncol.*, 22(6):1040–1044, 2004.

[69] Zena M Hira and Duncan F Gillies. A review of feature selection and feature extraction methods applied on microarray data. *Adv. Bioinf.*, 2015, 2015.

[70] Katherine A Hoadley, Christina Yau, Denise M Wolf, Andrew D Cherniack, David Tamborero, Sam Ng, Max DM Leiserson, Beifang Niu, Michael D McLellan, Vladislav Uzunangelov, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944, 2014.

[71] F. H. Hsu, E. Serpedin, T. H. Hsiao, A. J. Bishop, E. R. Dougherty, and Y. Chen. Reducing confounding and suppression effects in TCGA data: an integrated analysis of chemotherapy response in ovarian cancer. *BMC Genomics*, 13 Suppl 6:S13, 2012.

[72] T. Iba, J. Kigawa, Y. Kanamori, H. Itamochi, T. Oishi, M. Simada, K. Uegaki, J. Naniwa, and N. Terakawa. Expression of the c-myc gene as a predictor of chemotherapy response and a prognostic factor in patients with ovarian cancer. *Cancer Sci.*, 95(5):418–23, 2004.

[73] H Ishwaran, U B Kogalur, E H Blackstone, and M S Lauer. Random survival forests. *Ann. Appl. Stat.*, 2(3):841–860, 2008. URL `http://arXiv.org/abs/0811.1645v1`.

[74] A. A. Jazaeri, C. S. Awtrey, G. V. Chandramouli, Y. E. Chuang, J. Khan, C. Sotiriou, O. Aprelikova, C. J. Yee, K. K. Zorn, M. J. Birrer, J. C. Barrett, and J. Boyd. Gene expression profiles associated with response to chemotherapy in epithelial ovarian cancers. *Clin. Cancer Res.*, 11:6300–10, 2005.

[75] AP Jekunen, RD Christen, DR Shalinsky, and SB Howell. Synergistic interaction between cisplatin and taxol in human ovarian carcinoma cells in vitro. *Br. J. Cancer*, 69(2):299, 1994.

[76] Woojin Jeong, Sang-Bae Kim, Bo Hwa Sohn, Yun-Yong Park, EUN SUNG PARK, Sang Cheol Kim, Sung Soo Kim, Randy L Johnson, Michael Birrer, David SL Bowtell, et al. Activation of YAP1 is associated with poor prognosis and response to taxanes in ovarian cancer. *Anticancer Res.*, 34(2):811–817, 2014.

[77] S. Kamazawa, J. Kigawa, Y. Kanamori, H. Itamochi, S. Sato, T. Iba, and N. Terakawa. Multidrug resistance gene-1 is a useful predictor of paclitaxel-based chemotherapy for patients with ovarian cancer. *Gynecol. Oncol.*, 86: 171–6, 2002.

[78] J. Kang, A. D. D'Andrea, and D. Kozono. A DNA repair pathway-focused score for prediction of outcomes in ovarian cancer treated with platinum-based chemotherapy. *J. Natl. Cancer Inst.*, 104:670–81, 2012.

[79] Pratima S. Karnik, Sucheta Kulkarni, Xiao-Pu Liu, G. Thomas Budd, and Ronald M. Bukowski. Estrogen receptor mutations in tamoxifen-resistant breast cancer. *Cancer Res.*, 54(2):349–353, 1994. ISSN 0008-5472. URL `http://cancerres.aacrjournals.org/content/54/2/349`.

[80] Khalid S Khan, Regina Kunz, Jos Kleijnen, and Gerd Antes. Five steps to conducting a systematic review. *J. Royal Soc. Med.*, 96(3):118–121, 2003.

[81] Hugh Kikuchi, Anne Reiman, Jenifer Nyoni, Katherine Lloyd, Richard Savage, Tina Wotherspoon, Lisa Berry, David Snead, and Ian A Cree. Development and validation of a TaqMan Array for cancer mutation analysis. *Pathogenesis*, 3(1):1–8, 2016.

[82] Roger John Benjamin King and Mike W Robins. *Cancer Biology*. Pearson Education, Harlow, England, 2006.

[83] Antonis Koussounadis, Simon P Langdon, In Hwa Um, David J Harrison, and V Anne Smith. Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. *Sci. Rep.*, 5, 2015.

[84] Marcin Kowanetz and Napoleone Ferrara. Vascular endothelial growth factor signaling pathways: therapeutic perspective. *Clin. Cancer Res.*, 12(17):5018–5022, 2006.

[85] Eyal Krupka and Naftali Tishby. Incorporating prior knowledge on features into learning. In Marina Meila and Xiaotong Shen, editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 227–234, San Juan, Puerto Rico, 21–24 Mar 2007. PMLR. URL `http://proceedings.mlr.press/v2/krupka07a.html`.

[86] Max Kuhn. Caret package. *J. Stat. Softw.*, 28(5):1–26, 2008.

[87] Arnold J. Levine, Jill Bargonetti, Gareth L. Bond, Josephine Hoh, Kenan Onel, Michael Overholtzer, Archontoula Stoffel, Teresky, Christine A. Walsh, and Shengkan Jin. The p53 network. In Gerard P. Zambetti, editor, *The p53 Tumor Suppressor Pathway and Cancer*. Springer US, New York, 2005.

[88] Min Li, Jie Yin, Ning Mao, and Lingya Pan. Upregulation of phosphorylated cofilin 1 correlates with taxol resistance in human ovarian cancer in vitro and in vivo. *Oncol. Rep.*, 29(1):58–66, 2013.

[89] Crystal Linkletter, Derek Bingham, Nicholas Hengartner, David Higdon, and Q Ye Kenny. Variable selection for Gaussian process models in computer experiments. *Technometrics*, 2012.

[90] K. M. Lisowska, M. Olbryt, V. Dudaladava, J. Pamula-Pilat, K. Kujawa, E. Grzybowska, M. Jarzab, S. Student, I. K. Rzepecka, B. Jarzab, and J. Kupryjanczyk. Gene expression analysis in ovarian cancer - faults and hints from DNA microarray study. *Front. Oncol.*, 4:6, 2014.

[91] Y. Liu, Y. Sun, R. Broaddus, J. Liu, A. K. Sood, I. Shmulevich, and W. Zhang. Integrated analysis of gene expression and tumor nuclear image profiles associated with chemotherapy response in serous ovarian carcinoma. *PLoS One*, 7: e36383, 2012.

[92] Katherine L Lloyd, Ian A Cree, and Richard S Savage. Prediction of resistance to chemotherapy in ovarian cancer: a systematic review. *BMC Cancer*, 15(1): 117, 2015.

[93] Lawrence A. Loeb, Keith R. Loeb, and Jon P. Anderson. Multiple mutations and cancer. *Proc. Natl. Acad. Sci. U.S.A.*, 100(3):776–781, 2003. doi: 10.1073/pnas.0334858100. URL `http://www.pnas.org/content/100/3/776.abstract`.

[94] Christopher J. Lord and Alan Ashworth. The DNA damage response and cancer therapy. *Nature*, 481(7381):287–94, Jan 19 2012. URL `http://0-search.proquest.com.pugwash.lib.warwick.ac.uk/docview/921239589?accountid=14888`.

[95] Marcos Malumbres and Mariano Barbacid. RAS oncogenes: the first 30 years. *Nat. Rev. Cancer*, 3:459–465, June 2003. URL `http://dx.doi.org/10.1038/nrc1097`.

[96] Lainie P Martin, Thomas C Hamilton, and Russell J Schilder. Platinum resistance: the role of DNA repair pathways. *Clin. Cancer Res.*, 14(5):1291–1295, 2008.

[97] Andriy Marusyk and Kornelia Polyak. Tumor heterogeneity: causes and consequences. *Biochim. Biophys. Acta*, 1805:105–117, 2010. URL `https://doi.org/10.1016/j.bbcan.2009.11.002`.

[98] N. Matsumura, Z. Huang, T. Baba, P. S. Lee, J. C. Barnett, S. Mori, J. T. Chang, W. L. Kuo, A. H. Gusberg, R. S. Whitaker, J. W. Gray, S. Fujii, A. Berchuck, and S. K. Murphy. Yin yang 1 modulates taxane response in epithelial ovarian cancer. *Mol. Cancer Res.*, 7:210–20, 2009.

[99] Peter McCullagh. What is a statistical model? *Ann. Statist.*, 30(5):1225–1310, 10 2002. doi: 10.1214/aos/1035844977. URL `http://dx.doi.org/10.1214/aos/1035844977`.

[100] Gordon W McLean, Neil O Carragher, Egle Avizienyte, Jeff Evans, Valerie G Brunton, and Margaret C Frame. The role of focal-adhesion kinase in cancer—a new therapeutic opportunity. *Nat. Rev. Cancer*, 5(7):505–515, 2005.

[101] Roger McLendon, Allan Friedman, Darrell Bigner, Erwin G Van Meir, Daniel J Brat, Gena M Mastrogianakis, Jeffrey J Olson, Tom Mikkelsen, Norman Lehman, Ken Aldape, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008.

[102] Fabiola Medeiros, C Ted Rigl, Glenda G Anderson, Shawn H Becker, and Kevin C Halling. Tissue handling for genome-wide expression analysis: a review of the issues, evidence, and opportunities. *Arch. Pathol. Lab. Med.*, 131 (12):1805–1816, 2007.

[103] M. Mendiola, J. Barriuso, A. Redondo, A. Marino-Enriquez, R. Madero, E. Espinosa, J. A. Vara, I. Sanchez-Navarro, G. Hernandez-Cortes, P. Zamora, E. Perez-Fernandez, M. Miguel-Martin, A. Suarez, J. Palacios, M. Gonzalez-Baron, and D. Hardisson. Angiogenesis-related gene expression profile with independent prognostic value in advanced ovarian carcinoma. *PLoS One*, 3 (12):e4051, 2008.

[104] Kevin P Murphy. *Machine Learning: a Probabilistic Perspective*. MIT press, Cambridge, MA, 2012.

[105] National Institute for Health and Care Excellence. Guidance on the use of paclitaxel in the treatment of ovarian cancer. (TA55), 2003.

[106] National Institute for Health and Care Excellence. Erlotinib for the first-line treatment of locally advanced or metastatic EGFR-TK mutation-positive non-small-cell lung cancer. (TA258), 2012.

[107] National Institute for Health and Care Excellence. EGFR-TK mutation testing in adults with locally advanced or metastatic non-small-cell lung cancer. (DG9), 2013.

[108] National Institute for Health and Care Excellence. Afatinib for treating epidermal growth factor receptor mutation-positive locally advanced or metastatic non-small-cell lung cancer. (TA310), 2014.

[109] National Institute for Health and Care Excellence. Dabrafenib for treating unresectable or metastatic BRAF V600 mutation-positive melanoma. (TA321), 2014.

[110] National Institute for Health and Care Excellence. Vemurafenib for treating locally advanced or metastatic BRAF V600 mutation-positive malignant melanoma. (TA269), 2015.

[111] National Institute for Health and Care Excellence. Osimertinib for treating locally advanced or metastatic EGFR T790M mutation-positive non-small-cell lung cancer. (TA416), 2016.

[112] National Institute for Health and Care Excellence. Advanced breast cancer: diagnosis and treatment. (CG81), 2017.

[113] Daniel Navarro. *Learning statistics with R: A tutorial for psychology students and other beginners. (Version 0.5)*. University of Adelaide, Adelaide, Australia, 2015. URL `http://ua.edu.au/ccs/teaching/lsr`. R package version 0.5.

[114] John A Nelder and Roger Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.

[115] W. Netinatsunthorn, J. Hanprasertpong, C. Dechsukhum, R. Leetanaporn, and A. Geater. WT1 gene expression as a prognostic marker in advanced serous epithelial ovarian carcinoma: an immunohistochemical study. *BMC Cancer*, 6: 90, 2006.

[116] Cancer Genome Atlas Research Network et al. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, 2011.

[117] Cancer Genome Atlas Research Network et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–525, 2012.

[118] Cancer Genome Atlas Research Network et al. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499(7456):43–49, 2013.

[119] Ana S Neumann, Erich M Sturgis, and Qingyi Wei. Nucleotide excision repair as a marker for susceptibility to tobacco-related cancers: A review of molecular epidemiological studies. *Mol. Carcinog.*, 42(2):65–92, 2005.

[120] E. Obermayr, D. C. Castillo-Tong, D. Pils, P. Speiser, I. Braicu, T. Van Gorp, S. Mahner, J. Sehouli, I. Vergote, and R. Zeillinger. Molecular characterization of circulating tumor cells in patients with ovarian cancer improves their prognostic significance – a study of the OVCAD consortium. *Gynecol. Oncol.*, 128(1):15–21, 2013.

[121] CNAM Oldenhuis, SF Oosting, JA Gietema, and EGE De Vries. Prognostic versus predictive value of biomarkers in oncology. *Eur. J. Cancer*, 44(7): 946–953, 2008.

[122] Karim Pichara and Alvaro Soto. Local feature selection using Gaussian process regression. *Intelligent Data Analysis*, 18(3):319–336, 2014.

[123] Juho Piironen and Aki Vehtari. Projection predictive input variable selection for Gaussian process models. *arXiv preprint arXiv:1510.04813*, 2015.

[124] Helena Pópulo, José Manuel Lopes, and Paula Soares. The mTOR signalling pathway in human cancer. *Int. J. Mol. Sci.*, 13(2):1886–1918, 2012.

[125] Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *J. Mach. Learn. Res.*, 6(Dec): 1939–1959, 2005.

[126] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `https://www.R-project.org/`.

[127] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, Massachusetts, 2006.

[128] M. R. Raspollini, G. Amunni, A. Villanucci, V. Boddi, and G. L. Taddei. Increased cyclooxygenase-2 (COX-2) and P-glycoprotein-170 (MDR1) expression is associated with chemotherapy resistance and poor prognosis. analysis in ovarian carcinoma patients with low and high survival. *Int. J. Gynecol. Cancer*, 15:255–60, 2005.

[129] Matilda Rentoft, Philip John Coates, Göran Laurell, and Karin Nylander. Transcriptional profiling of formalin fixed paraffin embedded tissue: pitfalls and recommendations for identifying biologically relevant changes. *PloS One*, 7(4):e35276, 2012.

[130] G Ridgeway. Generalized boosted models: A guide to the gbm package, 2007.

[131] D. M. Roque, N. Buza, M. Glasgow, S. Bellone, I. Bortolomai, S. Gasparrini, E. Cocco, E. Ratner, D. A. Silasi, M. Azodi, T. J. Rutherford, P. E. Schwartz, and A. D. Santin. Class III beta-tubulin overexpression within the tumor microenvironment is a prognostic biomarker for poor overall survival in ovarian cancer patients treated with neoadjuvant carboplatin/paclitaxel. *Clin. Exp. Metastasis*, 31(1):101–10, 2014.

[132] Douglas T. Ross, Uwe Scherf, Michael B. Eisen, Charles M. Perou, Christian Rees, Paul Spellman, Vishwanath Iyer, Stefanie S. Jeffrey, de R. Van, Mark Waltham, Alexander Pergamenschikov, Jeffrey C.F. Lee, Deval Lashkari, and Dari Shalon. Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, 24(3):227–35, 03 2000. URL `http://0-search.proquest.com.pugwash.lib.warwick.ac.uk/docview/222642048?accountid=14888`.

[133] Patrick Royston and Douglas G Altman. External validation of a Cox prognostic model: principles and methods. *BMC Med. Res. Methodol.*, 13(1):33, 2013.

[134] R. Sabatier, P. Finetti, J. Bonensea, J. Jacquemier, J. Adelaide, E. Lambaudie, P. Viens, D. Birnbaum, and F. Bertucci. A seven-gene prognostic model for platinum-treated ovarian carcinomas. *Br. J. Cancer*, 105:304–11, 2011.

[135] Terrance Savitsky and Marina Vannucci. Spiked Dirichlet process priors for Gaussian process models. *Journal of Probability and Statistics*, 2010, 2010.

[136] Terrance Savitsky, Marina Vannucci, and Naijun Sha. Variable selection for nonparametric Gaussian process priors: Models and computational strategies. *Stat. Sci.*, 26(1):130, 2011.

[137] Shlomo S Sawilowsky. New effect size rules of thumb. *J. Mod. Appl. Stat. Methods*, 8(2):26, 2009.

[138] Robert E Schapire and Yoav Freund. *Boosting: Foundations and Algorithms*. MIT press, Cambridge, Massachusetts, 2012.

[139] M. P. Schlumbrecht, S. S. Xie, G. L. Shipley, D. L. Urbauer, and R. R. Broaddus. Molecular clustering based on ERα and EIG121 predicts survival in high-grade serous carcinoma of the ovary/peritoneum. *Mod. Pathol.*, 24:453–62, 2011.

[140] Markus S Schröder, Aedín C Culhane, John Quackenbush, and Benjamin Haibe-Kains. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics*, 27(22):3206–3208, 2011.

[141] M. Schwede, D. Spentzos, S. Bentink, O. Hofmann, B. Haibe-Kains, D. Harrington, J. Quackenbush, and A. C. Culhane. Stem cell-like gene expression in ovarian cancer predicts type II subtype and prognosis. *PLoS One*, 8:e57799, 2013.

[142] Z. E. Selvanayagam, T. H. Cheung, N. Wei, R. Vittal, K. W. Lo, W. Yeo, T. Kita, R. Ravatn, T. K. Chung, Y. F. Wong, and K. V. Chin. Prediction of chemotherapeutic response in ovarian cancer with DNA microarray expression profiling. *Cancer Genet. Cytogenet.*, 154:63–6, 2004.

[143] Alex Sigal and Varda Rotter. Oncogenic mutations of the p53 tumor suppressor: The demons of the guardian of the genome. *Cancer Res.*, 60(24): 6788–6793, 2000. ISSN 0008-5472. URL `http://cancerres.aacrjournals.org/content/60/24/6788`.

[144] Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.*, 39(5):1–13, 2011. URL `http://www.jstatsoft.org/v39/i05/`.

[145] I. Skirnisdottir and T. Seidal. The apoptosis regulators p53, bax and PUMA: Relationship and impact on outcome in early stage (FIGO I-II) ovarian carcinoma after post-surgical taxane-based treatment. *Oncol. Rep.*, 27(3):741–7, 2012.

[146] Elzbieta A Slodkowska and Jeffrey S Ross. MammaPrint$^{\text{TM}}$ 70-gene signature: another milestone in personalized medical care for breast cancer patients. *Expert Rev. Mol. Diagn.*, 9(5):417–422, 2009.

[147] D. Spentzos, D. A. Levine, M. F. Ramoni, M. Joseph, X. Gu, J. Boyd, T. A. Libermann, and S. A. Cannistra. Gene expression signature with independent prognostic significance in epithelial ovarian cancer. *J. Clin. Oncol.*, 22:4700–10, 2004.

[148] D. Spentzos, D. A. Levine, S. Kolia, H. Otu, J. Boyd, T. A. Libermann, and S. A. Cannistra. Unique gene expression profile based on pathologic response in epithelial ovarian cancer. *J. Clin. Oncol.*, 23:7911–8, 2005.

[149] Harald Steck, Balaji Krishnapuram, Cary Dehing-oberije, Philippe Lambin, and Vikas C Raykar. On ranking in survival analysis: Bounds on the concordance index. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1209–1216. Curran Associates, Inc., New York, 2008. URL `http://papers.nips.cc/paper/3375-on-ranking-in-survival-analysis-bounds-on-the-concordance-index.pdf`.

[150] Masashi Takano, Hiroshi Tsuda, and Toru Sugiyama. Clear cell carcinoma of the ovary: is there a role of histology-specific treatment. *J. Exp. Clin. Cancer Res.*, 31(53):35, 2012.

[151] Fei Tan, Carol J. Thiele, and Li Zhijie. Neurotrophin signaling in cancer. In Richard M. Kostrzewa, editor, *Handbook of Neurotoxicity*. Springer New York, New York, 2014.

[152] G Tapia and I Diaz-Padilla. Molecular mechanisms of platinum resistance in ovarian cancer. In I Diaz-Padilla, editor, *Ovarian Cancer - A Clinical and Translational Update*. InTech, London, England, 2013.

[153] Zivana Tezak, Marina V Kondratovich, and Elizabeth Mansfield. US FDA and personalized medicine: in vitro diagnostic regulatory perspective. *Pers. Med.*, 7(5):517–530, 2010.

[154] Christina Therkildsen, Troels K Bergmann, Tine Henrichsen-Schnack, Steen Ladelund, and Mef Nilbert. The predictive value of KRAS, NRAS, BRAF, PIK3CA and PTEN for anti-EGFR treatment in metastatic colorectal cancer: A systematic review and meta-analysis. *Acta Oncol.*, 53(7):852–864, 2014.

[155] Terry M Therneau and Patricia M Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000. ISBN 0-387-98784-3.

[156] Richard W Tothill, Anna V Tinker, Joshy George, Robert Brown, Stephen B Fox, Stephen Lade, Daryl S Johnson, Melanie K Trivett, Dariush Etemadmoghadam, Bianca Locandro, et al. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin. Cancer Res.*, 14 (16):5198–5208, 2008.

[157] W N Venables and B D Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL `http://www.stats.ox.ac.uk/pub/MASS4`. ISBN 0-387-95457-0.

[158] R. G. Verhaak, P. Tamayo, J. Y. Yang, D. Hubbard, H. Zhang, C. J. Creighton, S. Fereday, M. Lawrence, S. L. Carter, C. H. Mermel, A. D. Kostic, D. Etemadmoghadam, G. Saksena, K. Cibulskis, S. Duraisamy, K. Levanon, C. Sougnez, A. Tsherniak, S. Gomez, R. Onofrio, S. Gabriel, L. Chin, N. Zhang, P. T. Spellman, Y. Zhang, R. Akbani, K. A. Hoadley, A. Kahn, M. Kobel, D. Huntsman, R. A. Soslow, A. Defazio, M. J. Birrer, J. W. Gray, J. N. Weinstein, D. D. Bowtell, R. Drapkin, J. P. Mesirov, G. Getz, D. A. Levine, and M. Meyerson. Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. *J. Clin. Invest.*, 123(1):517–25, 2013.

[159] Bert Vogelstein and Kenneth W. Kinzler. Cancer genes and the pathways they control. *Nature Med.*, 10(8):789–99, 08 2004. URL `http://0-search.proquest.com.pugwash.lib.warwick.ac.uk/docview/223112178?accountid=14888`.

[160] F Randy Vogenberg. Predictive and prognostic models: implications for healthcare decision-making in a modern recession. *Am. Health Drug Benefits*, 2(6):218, 2009.

[161] U. Vogt, B. Falkiewicz, K. Bielawski, U. Bosse, and C. M. Schlotter. Relationship of c-myc and erbB oncogene family gene aberrations and other selected factors to ex vivo chemosensitivity of ovarian cancer in the modified ATP-chemosensitivity assay. *Acta Biochim. Pol.*, 47(1):157–64, 2000.

[162] Xin Wang, Camille Terfve, John C. Rose, and Florian Markowetz. HTSanalyzeR: an R/Bioconductor package for integrated network analysis of high-throughput screens. *Bioinformatics*, 27(6):879, 2011.

[163] LJ Wei. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Stat. Med.*, 11(14-15):1871–1879, 1992.

[164] Penny F Whiting, Anne WS Rutjes, Marie E Westwood, Susan Mallett, Jonathan J Deeks, Johannes B Reitsma, Mariska MG Leeflang, Jonathan AC Sterne, and Patrick MM Bossuyt. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann. Intern. Med.*, 155(8):529–536, 2011.

[165] P. D. Williams, S. Cheon, D. M. Havaleshko, H. Jeong, F. Cheng, D. Theodorescu, and J. K. Lee. Concordant gene expression signatures predict clinical outcomes of cancer patients undergoing systemic therapy. *Cancer Res.*, 69: 8302–9, 2009.

[166] William E Winter, G Larry Maxwell, Chunqiao Tian, Jay W Carlson, Robert F Ozols, Peter G Rose, Maurie Markman, Deborah K Armstrong, Franco Muggia, and William P McGuire. Prognostic factors for stage III epithelial ovarian cancer: a Gynecologic Oncology Group study. *J. Clin. Oncol.*, 25(24):3621–3627, 2007.

[167] Gordon C Wishart, Elizabeth M Azzato, David C Greenberg, Jem Rashbass, Olive Kearins, Gill Lawrence, Carlos Caldas, and Paul DP Pharoah. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Res.*, 12(1):R1, 2010.

[168] Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2017. URL `http://yihui.name/knitr/`. R package version 1.16.

[169] X. Yan, J. Yin, H. Yao, N. Mao, Y. Yang, and L. Pan. Increased expression of annexin A3 is a mechanism of platinum resistance in ovarian cancer. *Cancer Res.*, 70:1616–24, 2010.

[170] Kosuke Yoshihara, Atsushi Tajima, Tetsuro Yahata, Shoji Kodama, Hiroyuki Fujiwara, Mitsuaki Suzuki, Yoshitaka Onishi, Masayuki Hatae, Kazunobu Sueyoshi, Hisaya Fujiwara, et al. Gene expression profile for predicting survival in advanced-stage serous ovarian cancer across two independent datasets. *PloS One*, 5(3):e9615, 2010.

[171] Yuan Yuan, Eliezer M Van Allen, Larsson Omberg, Nikhil Wagle, Ali Amin-Mansour, Artem Sokolov, Lauren A Byers, Yanxun Xu, Kenneth R Hess, Lixia Diao, et al. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat. Biotechnol.*, 32(7):644–652, 2014.

[172] Hang Zhou and D. Suter. Improving Gaussian processes classification by spectral data reorganizing. In *2008 19th International Conference on Pattern Recognition*, pages 1–4, Dec 2008. doi: 10.1109/ICPR.2008.4761790.